Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079504

REVIEW

The Semantic Design Space of Retrieval-Augmented Recommender Systems: A Systematic Review of LLM-Based Approaches

Minhyeok Choi[1], **Imran Ahsan**[2], **Hyunwook Yu**[1], **Taeyoung Choe**[1] **and Mucheol Kim**[1,*]

[1]Department of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea
[2]Department of Smart Cities, Chung-Ang University, Seoul, Republic of Korea

*Corresponding Author: Mucheol Kim. Email: kimm@cau.ac.kr

Received: 22 January 2026; Accepted: 26 March 2026; Published: 08 May 2026

ABSTRACT: Large language models (LLMs) are increasingly integrated into recommender systems to support semantic reasoning, natural language understanding, and user-adaptive personalization. However, their reliance on static parametric knowledge and fixed representations limits robustness in dynamic environments, particularly under long-tail and cold-start conditions. Retrieval-augmented architectures have emerged to address these limitations by grounding LLMs in external, non-parametric knowledge sources. This systematic literature review synthesizes 138 peer-reviewed studies published between 2023 and 2025 in conferences and journals, focusing on retrieval-augmented and LLM-enhanced recommendation. We analyze these works through a three-dimensional framework covering: (i) domain application, (ii) semantic feature and representation design, and (iii) algorithmic strategies for retrieval and personalization. The review shows that current research is concentrated in general recommendation and information retrieval, that similarity/retrieval, user-item interaction, and textual content signals dominate semantic modeling, and that LLM and BERT-style encoders form the primary representation backbones, while graph-based, multimodal, and hybrid approaches remain comparatively underexplored. Algorithmically, most systems adopt generic LLM-centric modeling with limited use of retrieval optimization, reinforcement learning, or structure-aware strategies, and only sporadic attention to explicit cold-start, hallucination, and robustness treatment. By mapping co-occurrence patterns between domains, semantic features, representation choices, and strategy families, this review identifies concrete gaps and transfer opportunities for future work on retrieval-augmented recommendation and provides a structured reference for designing more context-aware, explainable, and data-efficient LLM-based recommender systems.

KEYWORDS: Large language models (LLMs); recommender system; retrieval-augmented generation (RAG); semantic features

1 Introduction

Recent Large language models (LLMs) [1,2] and recommender systems show outstanding performances across diverse tasks, including Question Answer (QA) tasks [3], information retrieval [4,5] and cold-start recommendation [6]. In contrast to traditional non-LLM-based recommender systems [7,8], current approaches increasingly integrate LLMs into the recommendation pipeline to enhance personalization and dynamic semantic understanding [9–13]. LLMs provide a powerful framework for capturing deep contextual representations from item descriptions and user behavior [14–16], enabling richer semantic modelling within modern recommender systems. However, even though there are various approaches dealing with LLMs-based recommender systems, they rely on parametric memory [17,18] from internal

knowledge bases, such as LLMs itself or user-item interactions. Therefore, they are susceptible to hallucination [19,20] when their internal knowledge deviates from real-time behavioral data. In the presence of hallucination, the model generates responses that are factually inaccurate with the underlying evidence. Static LLM and recommendation architectures cause a "semantic bottleneck" by failing to leverage fine-grained contextual signals in the personalization LLM prompt [21] and LLMs-based recommender systems [22]. To address these problems, retrieval-augmented LLMs, which incorporate non-parametric knowledge into LLMs and recommender systems gain broad adoption in recent studies [11,22,23]. In particular, these models query external knowledge bases—extensive repositories spanning diverse subjects and domains—to retrieve information relevant to the current input. The retrieved content is then integrated into the LLMs and recommender systems, allowing them to maintain accuracy and remain aligned up to date world knowledge [24]. Previous studies propose semantic-alignment that fuse LLMs with search-based technologies to better capture contextual signals [11,25–27]. They aim to compensate for the limitations of LLMs and recommender systems in accessing real-time and task-specific information. Retrieval-augmented generation (RAG), along with semantic retrieval, emerge as two prominent strategies. These methods help bridge the gap between LLM priors and real-world user intent by retrieving semantic features [11].

Conventional recommendation struggle with data sparsity, cold-start scenarios, and limited semantic understanding, which restrict their ability to model evolving user intent. Although LLMs-based recommenders provide richer reasoning and contextual modeling, they still rely on static parametric memory and thus suffer from hallucination and misalignment with real-time behavioral signals. These shared limitations highlight the need for retrieval-augmented architectures that can supply up-to-date, fine-grained external knowledge to both paradigms depicted in Table 1. This gap motivates the need for a systematic synthesis that organizes recent advances and clarifies the role of retrieval in building more reliable and context-aware recommendation models. This paper conducts a systematic literature review (SLR) of LLMs, recommender systems with a particular focus on retrieval-enhanced architectures. Based on three categories – domain characteristics, semantic search, and algorithmic strategies–our classification framework is applied to a curated dataset of 138 recent studies presented at top academic journals and conferences. The analysis highlights emerging trends in representation learning, semantic feature integration. This taxonomy enables the identification of existing research gaps and suggests potential directions for advancing future recommendation and LLMs in terms of RAG.

- A systematic literature review of 138 retrieval-augmented LLMs and recommender system studies published in top venues from 2023 to 2025 provides a quantitative analysis of current practices in semantic feature design, representation strategies.
- A classification framework with categories covering domain applications, semantic retrieval, and algorithmic strategies enables consistent cross-domain comparison and supports systematic benchmarking.
- This analysis not only identifies challenges that remain underexplored cold-start scenarios, but also outlines future research directions for advancing context-aware recommendation systems.

Table 1: Limitations of conventional vs. LLM-based recommender systems and the role of RAG.

Research Domain	Recent Limitations	Role of RAG
LLM-based recommendation	Hallucination, missing factual evidence	Grounding external knowledge, improving factual consistency
Conventional recommendation	Cold-start, long-tail, sparse interactions	Semantic retrieval, external signals to supplement sparse user-item data

2 Related Review Studies

A growing body of survey work examines LLMs and RAG from different perspectives. However, as summarized in Table 2, existing reviews remain only partially aligned with the requirements of LLM-based recommender systems. Most surveys emphasize high-level aspects such as hallucination mitigation, factual grounding, or generic retrieval architectures, but they rarely analyze the semantic features and user–item signals that drive personalized recommendation, nor do they systematically connect these signals to algorithmic strategies and domain-specific challenges. Arslan et al. [1] provide a comprehensive overview of RAG integrated with LLMs, categorizing more than 50 studies by task and discipline and showing that current RAG research is concentrated in QA, biomedical, and software-development applications. Fan et al. [28] and Church et al. [29] offer broader conceptual treatments of RAG as a mechanism for supplementing LLMs with external knowledge, organizing prior work by architectural pattern, retrieval strategy, and deployment scenario, and highlighting issues such as retrieval latency, evaluation inconsistency, and robustness. Zhao et al. [30], Swacha and Gracel [31], and Huang and Huang [32] similarly survey retrieval-augmented text generation pipelines, with taxonomies over retrieval methods, knowledge integration strategies, and generation models. Other surveys focus on more specific application settings. Li et al. [33] review RAG in educational contexts, emphasizing curated knowledge sources and reliability in feedback and recommendation tasks. Hu and Lu [34] broaden the scope beyond generation by introducing Retrieval-augmented Understanding (RAU) and a unified view of Retrieval-augmented Language Models (RALMs), covering both NLG and NLU tasks and discussing interaction patterns between retrievers and language models.

Table 2: Comparative analysis of recent surveys and our approach.

Survey	Year	Classification by Deep Feature Category	Domain Details	Open Challenges	Semantic Features
Arslan et al. [1]	2023–2024	✓	✓	×	✓
Fan et al. [28]	2021–2024	×	✓	×	✓
Church et al. [29]	2023–2024	×	✓	×	✓
Zhao et al. [30]	2021–2024	×	✓	✓	✓
Swacha and Gracel [31]	2021–2025	×	✓	×	×
Huang and Huang [32]	2020–2024	×	×	✓	✓
Li et al. [33]	2019–2025	×	✓	×	✓
Hu and Lu [34]	2016–2024	×	✓	✓	✓
Proposed survey	2023–2025	✓	✓	✓	✓

Collectively, these surveys establish RAG as a promising paradigm for knowledge-intensive LLM applications, and they identify important open issues around retrieval quality, scalability, domain adaptation, and robustness. Nevertheless, Table 2 shows that prior work only partially addresses the dimensions that are critical for recommendation. Most reviews do not classify systems by deep feature categories or domain-specific semantic signals, and only a subset explicitly discusses open challenges related to user–item interactions, long-tail exposure, or adaptive personalization. In particular, none of the existing surveys jointly analyze (i) the types of semantic features exploited for retrieval and representation, (ii) the recommendation domains and data characteristics in which these features appear, and (iii) the algorithmic strategies used to incorporate them into RAG-enhanced recommendation pipelines. In contrast, our survey introduces a unified taxonomy that integrates domain characteristics, semantic feature representations, and algorithmic

strategies for RAG-based recommendation. We systematically categorize the semantic features used in recent LLM-based recommender systems, examine their functional roles across retrieval and generation stages, and relate them to core recommendation challenges such as cold start, long-tail recommendation, and dynamic preference modeling. This domain-focused, feature-oriented, and algorithmically structured perspective complements existing RAG surveys and provides a sharper lens for understanding and advancing personalized RAG systems.

3 Data Extraction and Integration

During the data extraction phase, we extract and code each included study using a standardized extraction form aligned with the three analytical categories as shown in Fig. 1 (domain characteristics, semantic features/representations, and algorithmic strategies). We use a predefined codebook to ensure consistent assignment of labels across studies, and we record both (i) the extracted variables (e.g., domain label, semantic feature type, representation family, strategy family, evaluation setting) and (ii) supporting evidence (e.g., the relevant section/statement in the paper) needed for later verification. Appendix C reports the extraction template (field names) and the codebook (definitions and decision rules) that operationalize our framework and make the extraction procedure auditable.

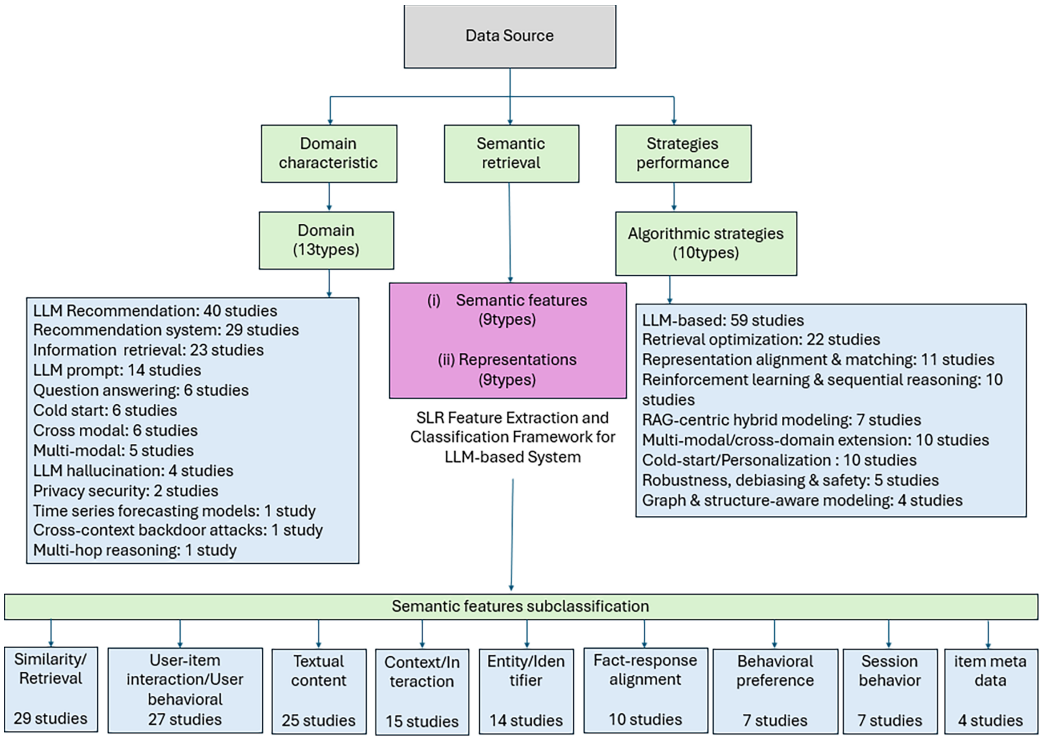

Figure 1: The figure summarizes the study selection process, core analytical categories, and extracted features across domains, semantic components, and algorithmic strategies in LLM-based recommender systems and conventional recommendation.

4 Research Methodology

This study aims to examine the design patterns, semantic structures, and algorithmic innovations that define retrieval-augmented LLM-based recommender systems. To guide this investigation, we formulate five research questions aligned with the core classification framework used in the analysis.

4.1 Research Questions

- **RQ1.** What are the primary application domains investigated in LLM-driven recommender systems that incorporate semantic retrieval techniques?
- **RQ2.** What semantic feature categories are employed in LLM-enhanced recommender systems, and how are they designed or claimed to address data sparsity, long-tail distributions, and cold-start scenarios?
- **RQ3.** What high-level algorithmic strategy families are employed, and how are these strategies used to support personalization, robustness, and long-tail recommendation?
- **RQ4.** How do different semantic feature types co-occur with different representation methods in retrieval-augmented LLM-based recommendation pipelines?
- **RQ5.** How do algorithmic strategy families distribute across the main application domains in retrieval-augmented LLM-based recommendation research?

4.2 Classification Framework

Our review protocol is structured around a three core classification framework that supports the analysis of each selected study. This framework ensures that all works are categorized consistently based on key functional components of LLM-based recommender systems.

1. Domain characteristic category: This category encompasses application domain-level characteristics such as long-tailed recommendation, cold-start scenarios, and domain heterogeneity, all of which directly influence a model's ability to generalize and adapt to diverse recommendation environments [35–37].
2. Semantic retrieval category: The semantic retrieval category examines how semantic features, contextual representations, and specialized retrieval policies are employed to enrich interaction-aware semantics and consequently elevate recommendation performance [38–40].
3. Strategies performance category: This category incorporates prompting methods, reasoning approaches, and learning paradigms aimed at enhancing robustness, personalization, and task performance, especially in long-tail and cold-start environment [6,41,42].

The strategy for data extraction and analysis is explicitly designed to align each identified study with the three core categories of our framework, allowing for a structured and comparable synthesis across the literature. The forthcoming section details the review protocol, which is firmly grounded in this classification scheme and guides the methodological flow of the study.

4.3 Preliminary Protocol Setup

We develop a review protocol informed by systematic literature review (SLR) methodology [43], incorporating the three principal categories of our analytical framework. This protocol specifies the inclusion criteria and search procedures, which are described in greater detail in the subsequent subsection.

4.4 Selection and Rejection Criteria

To maintain methodological rigor and ensure topical relevance, we define explicit inclusion and exclusion criteria that operationalize the scope of this review.

Inclusion criteria:

We include a study if it satisfies all of the following:

- **Topical relevance (LLM + recommendation + retrieval):** The title/abstract/full text indicates that the study (i) uses a large language model (LLM) or instruction-tuned/foundation language model, (ii) addresses a recommendation task or a recommender-system component (including conversational recommendation, sequential recommendation, or recommendation-related reranking/ranking), and (iii) incorporates retrieval augmentation or semantic retrieval (e.g., retrieval-augmented generation, dense/sparse retrieval, embedding-based retrieval, knowledge retrieval) as a functional component for grounding, personalization, ranking, or generation.
- **Publication window:** The study is published between 2023 and 2025 (inclusive), consistent with our goal of synthesizing recent developments in LLM-based retrieval-augmented recommendation.
- **Venue scope (scoping decision):** The study appears in one of the predefined peer-reviewed venues listed in Section 4.5 (seven conferences and six journals). We treat this venue restriction as a scoping decision for feasibility and focus; it does not imply that relevant work cannot appear in other outlets.
- **Peer-reviewed research article with accessible full text:** The study is a peer-reviewed conference/journal paper for which the full text is available. We include full papers and short papers when they provide sufficient methodological detail for coding and interpretation.
- **Empirical evidence with quantitative outcomes:** The study reports an empirical evaluation with quantitative outcomes relevant to recommendation and/or retrieval-supported recommendation (e.g., ranking metrics such as NDCG/Recall/HR@K, retrieval metrics, or quantitative user study/online evaluation results), sufficient to extract and compare evidence in our taxonomy.

Exclusion criteria:

We exclude a study if any of the following applies:

- The work does not involve an LLM (e.g., it only uses non-LLM encoders without an LLM component).
- The work is not related to recommendation (e.g., it focuses solely on generic QA/IR without a recommendation objective or recommendation-relevant pipeline component for our taxonomy).
- The work does not include retrieval augmentation or semantic retrieval as a functional component (e.g., purely parametric generation without retrieval or external evidence access).
- The work is not a peer-reviewed research paper (e.g., editorial, tutorial, keynote, extended abstract without methods), or the full text is not available.
- The work does not provide sufficient methodological detail or quantitative outcomes to support coding under our framework.

4.5 Search Process

To ensure both breadth and rigor within a feasible scope, we conduct a structured search over a predefined set of high-relevance peer-reviewed venues in NLP, information retrieval, recommender systems, and machine learning. Specifically, we cover conferences (EMNLP, ACL, NeurIPS, KDD, SIGIR, RecSys, and WSDM) and journals (ESWA, IPM, UMUAI, TORS, TKDE, and JAIR). We restrict the publication window to 2023–2025 to capture the most recent developments in retrieval-augmented and LLM-based recommendation research.

Information sources and access points.

We access each source through its official proceedings/journal portal (e.g., conference proceedings pages, journal issue listings, and/or venue-specific search interfaces when available). Appendix A reports the

access point used for each venue, the coverage years, and the last-accessed date recorded during our final verification pass (Table A1).

Query formulation and keyword filters.

We use a three-block concept formulation that reflects the overlap of (i) LLMs, (ii) recommendation, and (iii) retrieval augmentation. We instantiate this formulation through concrete keyword terms (e.g., "large language model"/LLM/GPT/foundation model; recommend*/recommender/conversational recommendation/sequential recommendation; retrieval-augmented generation/RAG/retrieval-augmented/semantic retrieval/dense retrieval/knowledge retrieval). We adapt query syntax to each venue portal when a search interface is available; otherwise, we apply the keyword filters during title/abstract screening after compiling venue-year listings. Appendix A reports the information sources, coverage years, keyword blocks, and portal-specific adaptation rules used for record identification Tables A2 and A3.

Limits and record management.

We limit consideration to peer-reviewed papers with accessible full text from the selected venues within 2023–2025. We compile all retrieved records into a master spreadsheet/bibliography file, and we check for duplicates using title and DOI (when available), merging any duplicates into a single record before screening.

Filtering workflow.

Fig. 2 summarizes the screening and eligibility workflow. In total, we retrieve 23,555 records from the selected venues. We then apply staged screening to identify the final set of 138 included studies:

1. **Title screening:** We screen titles for topical relevance under the criteria in Section 4.4 and exclude 21,890 records.
2. **Abstract screening:** We screen abstracts of the remaining 1665 records and exclude 1271 records that do not meet our eligibility requirements.
3. **Full-text eligibility assessment:** We assess 394 papers in full text and exclude 99 papers that fail at least one inclusion criterion.
4. **Detailed eligibility verification:** We perform detailed eligibility checks on 295 papers and exclude 157 papers due to insufficient methodological detail for coding and/or lack of quantitative outcomes aligned with our extraction framework.
5. **Final inclusion:** We include 138 studies in the final synthesis.

4.6 Study Selection Procedure

Two reviewers screen records in two stages: (i) title/abstract screening and (ii) full-text eligibility assessment. We perform screening independently and resolve disagreements through discussion until consensus. Because the review team consists of two screeners, we do not use a third adjudicator. During full-text assessment, we record an explicit exclusion reason code for each excluded paper (Appendix B), enabling an auditable eligibility trail.

4.7 Data Extraction and Coding Protocol

We operationalize the classification framework in Fig. 1 through a codebook that defines each label (domain category, semantic feature type, representation family, and strategy family) and specifies decision rules for ambiguous cases. We code domain labels and semantic/strategy categories as multi-label when a study clearly spans multiple categories (e.g., a conversational recommender that also proposes retrieval optimization). We record supporting evidence for each assigned label (e.g., the method description that indicates the retrieval component, the representation backbone, or the optimization strategy). Appendix C provides (i) the extraction form template (field names) and (ii) the full codebook used for labeling.

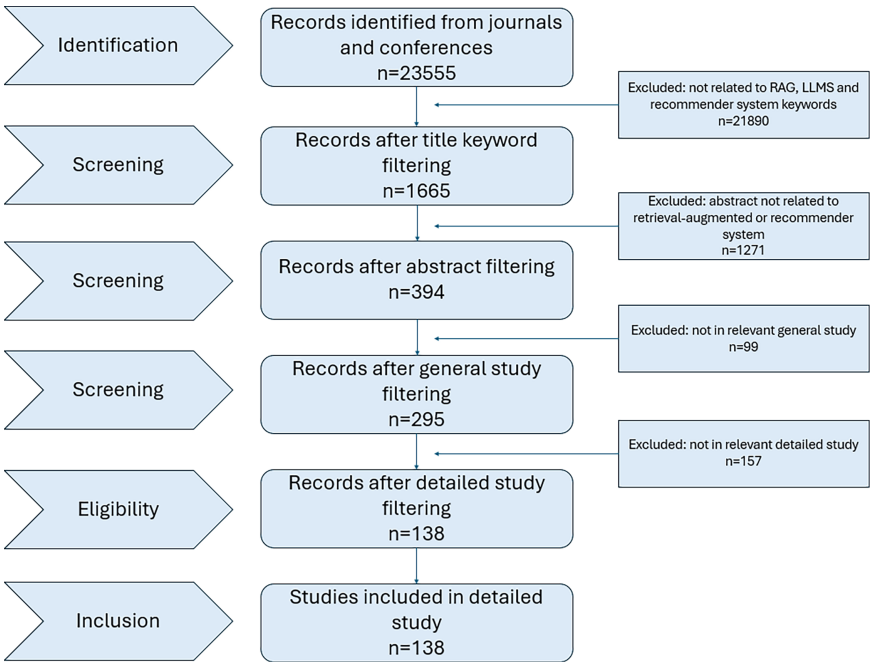

Figure 2: PRISMA-style flow diagram of the identification, screening, eligibility assessment, and inclusion process. The workflow narrows the initial pool of 23,555 records to 138 included studies under the predefined criteria in Section 4.4.

4.8 Quality Appraisal and Risk of Bias Assessment

Methodological transparency constitutes a central requirement for systematic literature review research. The review protocol follows a structured screening framework inspired by PRISMA principles. The process includes database identification, structured keyword construction, duplicate removal, title and abstract screening, followed by full-text eligibility assessment. Explicit inclusion and exclusion criteria guide the filtering procedure in order to reduce discretionary judgment during study selection.

Selection bias represents a primary methodological concern in emerging research domains. Inclusion criteria require explicit integration of RAG or LLM components, empirical validation, clear evaluation reporting, and sufficient methodological description to enable comparative analysis. Exclusion criteria remove purely conceptual discussions, non-empirical commentaries, inaccessible manuscripts, and studies lacking evaluative transparency. Such filtering enhances analytical coherence yet introduces potential structural bias, particularly in fast-evolving areas where workshop papers or preprints contribute substantially to technical developmemt. Broad keyword combinations alongside multi-database coverage mitigate this risk by expanding search sensitivity.

Quality appraisal operates through a structured evaluative lens. Each included study receives examination with respect to empirical grounding, dataset transparency, clarity of metric specification, reproducibility-related description, and experimental reporting consistency. The framework prioritizes minimal methodological adequacy rather than numerical scoring, thereby ensuring baseline research reliability across the corpus.

Quantitative descriptive synthesis strengthens analytical robustness. Metric frequency distribution, dataset category prevalence, task alignment patterns, and evaluation concentration effects receive systematic aggregation across the 138 systems. Such aggregation reveals structural regularities within the evaluation ecosystem while preserving the exploratory scope of the review.

5 Results

This section presents the key findings derived from the systematic review of the 138 studies that meet all inclusion criteria. Each selected work contributes to at least one of the core analytical dimensions—semantic retrieval mechanisms, domain/data characteristics, and algorithmic strategies for personalization—which collectively outline the methodological landscape of LLM-driven recommender systems. We organize the results according to RQ1–RQ5 to provide a systematic and analytically grounded understanding of methodological trends, conceptual developments, and open challenges in RAG-and LLM-enhanced recommender systems.

5.1 Application Domains

Domain distribution. Domain distribution refers to how the 138 included studies are spread across application areas that are relevant to LLM-based recommendation and closely related retrieval-augmented tasks. As shown in Table 3, we distinguish domains such as LLM Recommendation, conventional Recommendation system, Information retrieval, Question answering, Multi-modal and Cross-modal tasks, Cold start, LLM hallucination, LLM prompt, privacy and security, time series forecasting models, cross-context backdoor attacks, and Multi-hop reasoning. This taxonomy provides the contextual layer for interpreting how domain characteristics influence model design and evaluation.

LLM recommendation and conventional recommendation system. We identify **40 studies** in the *LLM Recommendation* category (for example, [9,44,45], etc.), making it the most prevalent domain. These works treat LLMs as core components for personalized information access, using them to interpret user intent [12], encode rich semantic signals, and generate recommendation outputs through natural-language reasoning and retrieval-guided inference [11]. Compared with conventional collaborative or content-based recommenders [23,46], these systems reposition language models as both semantic processors and decision-making engines. The category of *Recommendation system* domain contains **29 studies**. These works focus on predicting user preferences using collaborative filtering [47], content-based modeling [48], or hybrid approaches that primarily leverage historical interactions [49], often with fixed representation pipelines such as matrix factorization [50] or sequential neural models. LLMs may appear as auxiliary components (e.g., for feature construction or explanation), but the core modeling remains closer to traditional recommendation paradigms, with an emphasis on scalability, sparsity mitigation, and sequential preference modeling.

Information retrieval and question answering. *Information retrieval (IR)* accounts for **23 studies**. These works center on retrieving and ranking relevant documents, passages, or entities from large corpora, frequently serving as the retrieval backbone for RAG architectures. Advances focus on query representation [24], retrieval effectiveness [51,52], and corpus indexing [53], often in combination with QA-style evaluation. *Question answering (QA)* is explicitly targeted in **6 studies**. These papers concentrate on generating accurate, evidence-grounded answers to natural-language queries and are frequently used as benchmarks for evaluating how well RAG-enhanced systems align retrieved evidence with generated responses. While QA is not recommendation per se, methods developed in this domain (e.g., dense retrieval and evidence aggregation) are directly transferable to recommendation scenarios where explanations or justifications are required.

Prompt-centric and hallucination-focused studies. The *LLM prompt* domain comprises **14 studies**, where prompt design is treated as a primary control mechanism for steering LLM behaviour in recommendation or retrieval tasks. These works explore how prompt templates, instructions, and few-shot examples influence personalization, output format, and retrieval-augmented reasoning without additional parameter tuning [54,55]. *LLM hallucination* is the focus of **4 studies**. They investigate failure modes in which LLMs produce fluent but incorrect or unverifiable content, and they propose retrieval-augmented or verification

mechanisms to reduce hallucination. Although most experiments are conducted in QA or open-domain generation settings, the mitigation strategies are relevant to recommendation scenarios that require factual grounding (e.g., item attributes, reviews, or external knowledge).

Cold start, multi-modal, and cross-modal domains. Six studies address *cold-start* recommendation, where user or item representations are poorly developed due to sparse or non-existent interaction histories. These works typically emphasize constructing robust initial representations from content, side information, or synthetic interaction signals generated by LLMs, and they are particularly important for understanding how RAG can alleviate extreme sparsity [36]. The *multi-modal* and *cross-modal* domains together contribute **11 studies**. Five papers treat multi-modal recommendation or retrieval, integrating text, images, and structured attributes into joint representations for richer semantic modeling. Six papers focus on cross-modal tasks such as aligning text with images or other modalities, emphasizing inter-modal translation and semantic correspondence. These domains are of particular interest for recommendation contexts where items are described by heterogeneous content (e.g., products, multimedia) [56,57].

Other specialized domains. The remaining domains appear only in a small number of studies: *privacy and security* (2 studies [58,59]), *time series forecasting* (1 study [60]), *Cross-Context Backdoor Attacks* (1 study [61]), and *multi-hop reasoning* (1 study [62]). These works highlight emerging concerns such as secure retrieval and adversarial vulnerabilities, or extend retrieval-augmented modeling to temporal and graph-based reasoning tasks, but they are still peripheral in the current RAG-for-recommendation landscape.

The distribution of domains in Table 3 indicates that retrieval-augmented and LLM-based approaches are predominantly evaluated in classical recommendation and information-retrieval settings, whereas cross-domain, multi-modal, and security-oriented scenarios receive only limited attention. This pattern is consistent with a field that is still consolidating basic architectures before moving into more demanding application contexts. At the same time, the absence of standardized reporting on user–item sparsity, long-tail behaviour, and domain-specific biases limits our ability to judge how robust these methods would be under more realistic and heterogeneous deployment conditions.

Table 3: Application domains of the retrieved RAG/LLM studies, with corresponding paper identifiers. Percentages are relative to the 138 papers for which a domain label is assigned.

Domain	Paper IDs	Number of Papers (N = 138)	Share of Domain-Labelled Papers (%)
Information retrieval	[4,5,24,51–53,63–79]	23	16.8
LLM recommendation	[3,9–13,15–17,20,22,27,37–41,44,45,80–105]	40	29.0
LLM hallucination	[19,20,106,107]	4	2.9
Recommendation system	[7,8,14,21,23,25,38,46–50,91,108–123]	29	21.0
Multi-modal	[56,124–127]	5	3.6
Question answering	[3,128–132]	6	4.4
Cold start	[6,35–37,42,133]	6	4.4
Cross modal	[26,57,134–137]	6	4.4
Privacy security	[58,59]	2	1.5
Time series forecasting models (TSFMs)	[60]	1	0.7

(Continued)

Table 3 (continued)

Domain	Paper IDs	Number of Papers (N = 138)	Share of Domain-Labelled Papers (%)
LLM prompt	[18,54,55,138–148]	14	10.1
Cross-Context Backdoor Attacks	[61]	1	0.7
Multi-hop reasoning	[62]	1	0.7

5.2 Semantic Techniques for Hallucination and Cold-Start Challenges

Semantic features. Semantic features capture high-level signals about users and items that go beyond raw interaction counts, enabling models to infer intent, contextual relevance, and preference structure under sparse or noisy conditions. In our corpus, we identify nine feature categories: user-item interaction/user-item behavioral signals, textual content, similarity/retrieval signals, behavioral preference, context/interaction, entity/identifier features, fact-response alignment, session behavior, and item metadata. These feature types are central to how recent systems are *designed or claimed* to address hallucination, long-tail exposure, and cold-start limitations, where classical collaborative filtering alone is insufficient [14,108].

Similarity/retrieval features are the most common category, appearing in **29 studies** (Table 4). These features encode relational semantics by measuring proximity between users, items, or textual contexts in lexical or embedding space [53]. Typical implementations rely on vector similarity (e.g., cosine or dot-product) and dense retrieval signals to identify evidence aligned with user intent [41,51]. By grounding decisions in retrieval-aware similarity structures, these features strengthen robustness under data sparsity and help align LLM reasoning with external knowledge, which is particularly important when recommending long-tail items or operating in cold-start regimes. **User–item interaction/user–item behavioral** features appear in **27 studies** and capture the traces left by historical engagements such as clicks, views, purchases, and ratings [14,47]. These signals reflect implicit preference dynamics over time and provide a behavior-grounded view of user interests that complements content-only semantics [10]. In LLM-based recommenders, interaction sequences are often used as semantic priors or prompts that guide preference reasoning and context-aware decision-making [6,17]. Their granularity is especially valuable in sparse environments, where a small number of interactions must be exploited carefully to model realistic decision trajectories.

Textual content, present in **25 studies**, captures the intrinsic semantics encoded in natural-language item descriptions such as titles, synopses, and reviews [108,117]. Unlike interaction features, which summarize user behaviour, textual content provides content-grounded evidence that is available even in the absence of rich interaction histories. It also differs from similarity/retrieval features, which focus on relational proximity rather than inherent linguistic meaning. For RAG-enhanced recommenders, textual content offers a stable semantic basis that allows LLMs to understand item characteristics and support cold-start reasoning when behavioural data are limited [141]. **Context/interaction** features occur in **15 studies** and model situational or session-level signals such as co-occurring actions, conversational turns, or short-term temporal dependencies [64,128]. These features characterize dynamic behavioural states that evolve during a user's engagement with the system [61,121], capturing immediate intent and contextual relevance that cannot be inferred from item text alone. In LLM-based settings, context/interaction features act as a real-time layer that complements static semantic content and helps adapt recommendations to recent behaviour.

Table 4: Semantic feature types used in the 138 retrieved studies, with corresponding paper identifiers.

Semantic Feature Type	Paper IDs	Number of Papers (N = 138)	Share of All Papers (%)
User-item interaction/ User-item behavioral	[4,6,10,14,15,17,18,23,37,38,40,42,46,47,49, 76,81,82,92,93,98,100–102,104,105,119]	27	19.6
Textual content	[11,12,16,24,26,45,48,59,63,65,71,72,74,78,83, 91,108,111,113,115,117,122,123,141,145]	25	18.1
Similarity/Retrieval	[13,25,41,44,51–53,55–58,66,75,77,79,84,106, 109,110,114,124,127,130,132,137,138,142,143,148]	29	21.0
Behavioral preference	[8,21,35,80,94,97,126]	7	5.1
Context/Interaction	[22,36,61,64,70,73,87,88,99,103,118,121,128, 131,146]	15	10.9
Entity/Identifier	[9,27,54,62,68,96,125,129,134– 136,139,144,147]	14	10.1
Fact-response alignment	[3,5,19,20,39,67,85,95,107,120]	10	7.1
Session behavior	[7,50,60,69,90,112,116]	7	5.1
Item metadata	[86,89,133,140]	4	2.9

Entity/identifier features are used in **14 studies** and denote discrete symbols that uniquely index users, items, or domain-specific objects [96,147]. They typically carry no intrinsic semantic meaning, but serve as stable anchors for associating structured attributes, relational metadata, or historical statistics with specific entities. In contrast to behavioural signals, identifiers capture object identity rather than user actions. They are often combined with other feature types to link raw IDs to richer semantic or structural information [27,129]. **Fact-response alignment** features appear in **10 studies** and measure how faithfully an LLM's generated output is grounded in retrieved or provided evidence [3,19]. These features evaluate whether claims in the response are supported, contradicted, or unverifiable with respect to the underlying evidence [67,120]. By emphasizing evidence-conditioned truthfulness, fact-response alignment provides a semantic constraint for diagnosing and mitigating hallucinations in retrieval-augmented generation [20,39].

Behavioral preference features are adopted in **7 studies** and aim to capture latent user inclination inferred from patterns in historical engagement [80,97]. Rather than representing individual events, these features summarize higher-level tendencies and context-dependent shifts in intent [21,126]. They provide an additional behavioural layer that can guide personalized reasoning, particularly when explicit feedback is scarce or noisy [94]. **Session behavior** features, also present in **7 studies**, focus on short-term patterns within a single interaction episode [7,90]. They encode high-resolution, momentary signals that reveal the user's current goal and situational context [60,112], supporting the modeling of intent drift and real-time adaptation in both traditional and LLM-based recommendation. Session-level features therefore complement long-term histories and static content features in settings such as conversational or sequential recommendation. **Item metadata** features appear in **4 studies** and refer to structured, non-textual attributes such as categories or platform-level descriptors [133,140]. These attributes provide concise, standardized signals about item functionality that are independent of user behaviour or learned embedding spaces. Item metadata offers a

stable scaffold for representation learning, assisting cold-start reasoning and hierarchical organization in LLM-based recommenders [89,140].

As summarized in Table 4, current systems are moving toward hybrid semantic spaces that combine interaction, content, retrieval, and contextual signals. Recommendation is no longer driven solely by co-occurrence patterns; instead, semantically rich evidence increasingly supports generalization under sparse, dynamic, and previously unseen conditions. This shift reinforces the role of semantic feature design as a core ingredient for long-tail and cold-start reasoning.

Representation methods. Semantic representation methods map heterogeneous raw signals into structured, machine-interpretable embeddings that support preference modeling, context reasoning, and cross-domain generalization [11,12,139]. In our corpus, we distinguish nine representation families: LLM-based encoders, BERT/Sentence-transformer models, classic distributional embeddings, multimodal representations, graph-based representations, sequential user representations, hybrid content-collaborative representations, tabular/attribute-based encodings, and structured reasoning representations.

LLM-based encoders are the most common representation family, used in **37 studies**. These models provide high-capacity semantic encoders that map text, metadata, and other structured signals into unified, context-rich representation spaces [64,99]. In contrast to traditional embedding models, LLMs generate deeply contextualized representations that can be adapted through prompting or instruction-tuning, making them attractive backbones for RAG-enhanced recommendation [95,130]. **BERT and Sentence-transformer** models form a second major family, appearing in **29 studies**. They project textual inputs into dense, context-aware embedding spaces that support retrieval and ranking [7,120]. Compared with LLM-scale encoders, these models offer a more efficient balance between semantic richness and computational cost, and they remain a widely used choice for dense retrieval and similarity-based recommendation. **Classic distributional embeddings** appear in **15 studies** and encode word semantics based on co-occurrence statistics in large corpora. Although they lack contextualization, their lightweight structure supports efficient retrieval and ranking, and they continue to serve as interpretable baselines for assessing newer contextual encoders.

Multimodal representations are used in **12 studies**. These models incorporate text, images, and other modalities into joint embedding spaces, enabling richer item semantics and grounded reasoning over visual and linguistic cues. They are particularly relevant for recommendation settings where items are described by heterogeneous content [124,135]. **Graph-based representations** present in **9 studies**, encode entities and their relational topology using graph neural networks or related techniques (like [61,105], etc.). These representations capture interaction patterns, neighborhood context, and higher-order structure in user-item graphs or knowledge graphs, supporting tasks such as relation-aware recommendation and robust retrieval. **Sequential user representations** appear in **9 studies**. They model the ordered sequence of past interactions using recurrent, attention-based, or transformer-style architectures, capturing temporal dependencies and intent shifts more explicitly than static embeddings [60,121]. **Hybrid content-collaborative representations** are adopted in **10 studies**. These models fuse content-derived semantics (e.g., text or images) with collaborative signals from user-item interactions, aiming to combine the strengths of content-based and collaborative filtering while mitigating their respective weaknesses, particularly under cold-start conditions [4,41]. **Tabular or attribute-based representations**, used in **9 studies**, encode entities using structured fields such as categorical attributes and numerical descriptors (like [63,107], etc.). They provide explicit, interpretable signals grounded in item and user properties and often serve as complementary features alongside learned embeddings. **Structured reasoning representations** are also present in **8 studies**. These approaches encode intermediate logical or symbolic states to support multi-step reasoning, improve interpretability, and offer controllable interfaces for integrating external knowledge or constraints [98,118].

Across these representation families, Table 5 indicates a clear tendency to rely on general-purpose transformer-based encoders, particularly LLMs and BERT-style models, while more specialized or lightweight representations remain comparatively underexplored. This pattern suggests an opportunity for future work to more systematically combine LLM-based encoders with task-specific, multimodal, or hybrid representations, especially in settings that demand fine-grained personalization, efficiency, and robustness under domain shift.

Table 5: Representation methods used to encode semantic features in the 138 studies, with corresponding paper identifiers.

Representation Family	Paper IDs	Number of Papers (N = 138)	Share of all Papers (%)
LLMs	[3,9,14–18,20,22,27,39,44,58,64,72,73,81–83,86–90,92,93,95,96,99,100,102,103,110,130,133,138,145]	37	26.8
BERT/Sentence-transformer	[5,7,12,19,21,36,37,52–54,66,67,74–78,80,84,94,97,104,120,123,129,139–141,148]	29	21.0
Classic distributional embeddings	[24–26,47,48,51,70,91,108,114,115,119,122,132,134]	15	10.9
Multimodal representations	[56,57,79,111,124–127,131,135–137]	12	8.7
Graph-based representations	[10,13,40,55,61,62,68,105,113]	9	6.5
Sequential user representations	[6,42,50,60,69,71,109,121,128]	9	6.5
Hybrid content–collaborative representations	[4,11,23,41,45,49,65,101,106,116]	10	7.3
Tabular/Attribute	[8,35,46,59,63,85,107,112,117]	9	6.5
Structured reasoning representations	[38,98,118,142–144,146,147]	8	5.8

5.3 Strategies Performance for Personalization and Robustness

Algorithmic strategies. Algorithmic strategies refer to the modeling approaches that guide, adapt, or optimize recommendation and retrieval in LLM-based systems. In our taxonomy, we distinguish nine major families: (i) LLMs-based strategies, (ii) retrieval optimization, (iii) representation alignment and matching, (iv) graph-and structure-aware modeling, (v) RAG-centric hybrid modeling, (vi) reinforcement learning and sequential reasoning, (vii) robustness, debiasing and safety, (viii) multi-modal and cross-domain extension, and (ix) cold-start/personalization strategies. Together, these families represent the core computational mechanisms through which recent work seeks to personalize recommendations, ground outputs in external evidence, and improve robustness.

Across the 138 studies, **LLMs-based strategies** constitute the most common family (Table 6). These approaches rely on pretrained LLMs as primary decision-makers, often with minimal architectural modification (for instance [14,27], etc.). LLMs are used for tasks such as preference elicitation, candidate filtering, policy planning, or multi-step reasoning, typically via prompting, in-context learning, or chain-of-thought style decision chains [73,146]. This dominance reflects the field's early emphasis on validating the utility

of general-purpose LLMs before introducing more specialized model components. **Retrieval optimization strategies** focus on improving the quality, relevance, and efficiency of retrieved evidence by refining scoring functions, selection criteria, and retrieval policies [71,142]. Methods in this family incorporate mechanisms such as semantic filtering, uncertainty-aware sampling, or adaptive retrieval thresholds to better align retrieved contexts with user intent and task constraints [75,129,138]. Retrieval optimization directly affects downstream reasoning and recommendation performance in RAG pipelines and appears in **22 studies. Multi-modal and cross-domain extension strategies** enrich LLM-based recommenders by either integrating additional modalities or transferring knowledge across domains. Multi-modal and cross-modal methods unify textual, visual, and structural signals to strengthen personalization in scenarios where non-textual evidence helps disambiguate user intent (for example, [126,127], etc.). Cross-domain approaches, by contrast, aim to reuse preferences, latent structures, or behavioral patterns across heterogeneous recommendation settings, thereby reducing dependence on dense domain-specific data and improving robustness in cold-start and long-tail conditions. In total, these strategies appear in **10 studies**.

Table 6: High-level algorithmic strategy families adopted in the 138 RAG/LLM-based systems, with corresponding paper identifiers.

Strategy Family	Paper IDs	Number of Papers (N = 138)	Share of All Papers (%)
Retrieval optimization	[24,25,50,52,55,56,60,63,65,68,70–72, 75,77,78,94,107,128,129,138,142]	22	15.9
Representation alignment & matching	[26,47,51,53,66,108–110,124,134,135]	11	8.0
Graph & structure-aware modeling	[57,61,105,113]	4	2.9
LLMs-based	[3,5–7,9–14,16–18,20–22,27,39–41,44,45, 59,64,73,74,76,80–83,85–89,93,95– 104,122,123,130,132,133,140,141,144–148]	59	40.6
RAG-centric hybrid modeling	[15,19,23,38,67,96,139]	7	6.5
Reinforcement learning & sequential reasoning	[8,35,36,46,62,69,90,92,106,121]	10	8.7
Robustness, debiasing & safety	[54,58,84,118,120]	5	3.6
Multi-modal & cross-domain extension	[42,48,79,111,125–127,131,136,137]	10	7.3
Cold-start/Personalization	[37,49,91,112,114–117,119,143]	10	6.5

Reinforcement learning (RL) strategies treat recommendation and retrieval as sequential decision-making problems. Rather than relying solely on static heuristics or supervised labels, RL-based methods learn policies that maximize long-term user utility by adaptively updating actions such as retrieval selection, ranking refinement, or prompt routing based on reward signals [69,90,121]. These strategies are particularly relevant in interactive or evolving environments where user feedback is incremental and delayed [35,92]. A subset of works combines RL with explicit sequential reasoning, constructing multi-step inference chains

that refine predictions based on intermediate results [46,92]. Together, RL and sequential reasoning appear in a relatively small but growing group of studies. **Graph-and structure-aware modeling strategies** exploit the relational topology of user-item interactions, knowledge graphs, or hierarchical content structures [113]. These methods use graph neural networks or relational reasoning modules to leverage node connectivity, neighborhood aggregation, and higher-order structural dependencies [57,105]. Such strategies are particularly useful for capturing complex preference patterns, propagating signals across sparse interaction graphs, and enhancing robustness to missing data. In our corpus, they remain comparatively less frequent but indicate an emerging interest in structure-aware RAG for recommendation. **Representation alignment and matching strategies** aim to reconcile heterogeneous embedding spaces—such as user, item, textual, and retrieved evidence representations—into a shared, semantically coherent latent space [110]. Typical techniques include contrastive learning, metric learning, projection heads, and cross-modal matching functions [66,135]. By explicitly modeling cross-space correspondence, these methods help bridge gaps between behaviour-driven and content-driven features, and support more accurate matching in retrieval and recommendation. This family appears in a moderate number of studies (such as [26,66], etc.).

RAG-centric hybrid modeling integrates retrieval-augmented generation with traditional recommendation architectures. In these systems, retrieved evidence and generative reasoning jointly influence ranking decisions or item selection [139]. Such designs allow LLMs to ground their reasoning in external knowledge while retaining domain-specific inductive biases encoded in classical recommenders, aiming to improve robustness, interpretability, and long-tail performance [15,19]. This family includes works that couple LLMs with collaborative filtering, graph-based models, or session-based recommenders (for example, [38,96], etc.). Finally, **robustness-focused and cold-start/personalization strategies** target specific challenges. Robustness, debiasing, and safety methods investigate issues such as adversarial attacks, exposure bias, or hallucination under noisy retrieval (like [84,120], etc.). Cold-start and personalization strategies explicitly address sparse or non-existent interaction histories, using semantic features, data augmentation, or synthetic interactions to bootstrap user and item representations (for example, [91,143], etc.). Although these families are smaller in absolute size, they are directly relevant for real-world deployment of RAG-enhanced recommenders, where safety and data sparsity are practical constraints. Table 6 shows that most current systems rely on relatively generic LLM-based strategies, with more specialized approaches—such as retrieval optimization, reinforcement learning, structure-aware modeling, and explicit robustness mechanisms—still emerging. Few studies provide systematic comparisons of these strategies under long-tail or cold-start conditions, leaving open questions about which mechanisms most effectively improve generalization in sparse, dynamic environments. Addressing this gap through controlled evaluations and combined strategy design remains an important direction for future work.

5.4 Feature–Representation Co-Occurrence

To address RQ4, cross-tabulates semantic feature types with the primary representation families adopted in the 138 studies. The matrix highlights how different signals are actually encoded in practice, moving beyond the one-dimensional counts reported in Tables 4 and 5. User–item interaction and user–item behavioral features are most frequently paired with LLM-based encoders (10 papers) and hybrid content–collaborative representations (4 papers), with additional use of graph-based and classic embeddings. This pattern suggests that fine-grained behavioural traces are often treated as high-value signals that warrant either powerful contextual encoders or structure-aware models, rather than being handled by lightweight representations alone. Textual content is distributed more evenly across representation families, with notable counts for BERT/Sentence-transformer (6 papers), classic distributional embeddings (6), and hybrid representations (5), alongside LLM-based encoders (4). This spread reflects the coexistence of several design

choices: some systems rely on general-purpose LLM encoders for content, while others retain more efficient BERT-style or classic embeddings for text, especially when retrieval efficiency is a concern.

Similarity/retrieval features show a distinct profile. They are strongly associated with BERT/Sentence-transformer encoders (7 papers) and multimodal representations (6), but also appear with LLMs, classic embeddings, and graph-based models. This confirms that dense retrieval pipelines remain a primary use case for BERT-style encoders, while multimodal retrieval architectures leverage cross-modal similarity signals to bridge text and non-text content. Context/interaction features and entity/identifier features are each realized through a diverse set of representations, with non-trivial counts across LLMs, BERT/ST, graph-based, and structured reasoning encoders. This diversity indicates that there is not yet a dominant design pattern for encoding situational and identity-level signals; instead, these features are integrated opportunistically depending on the surrounding architecture. In contrast, several feature types exhibit more focused representation choices. Behavioral preference is almost exclusively encoded with BERT/ST (4 papers) and, to a lesser extent, multimodal and tabular representations, suggesting that compact, contrastive-style encoders are often deemed sufficient for summarizing higher-level preference tendencies. Session behavior is strongly tied to sequential user representations (4 papers), which is consistent with its role in modeling short-term, temporally ordered interactions. Fact–response alignment features rely mainly on BERT/ST and LLM-based encoders, with some use of tabular and structured reasoning representations, reflecting their emphasis on evidence-conditioned text understanding. Item metadata, although less frequent overall, tends to be combined with LLM-based and hybrid representations, where it complements richer semantic embeddings.

The patterns in Table 7 indicate that interaction-and session-level signals are rarely encoded with simple, static embeddings and instead gravitate toward LLM-, sequential-, or graph-based models, whereas textual and similarity features are distributed across both heavyweight and lightweight encoders. This asymmetry suggests that behavioural and contextual information are treated as harder modelling targets that benefit from more expressive or structure-aware representations, while content and similarity signals are often delegated to established text-embedding pipelines. From a design perspective, it would be valuable for future RAG-based recommenders to explore more deliberate pairings between semantic features and representations, particularly underexplored combinations such as session behaviour with multimodal encoders or fact–response alignment with graph-based models.

Table 7: Co-occurrence of semantic feature types and representation families. Each cell reports the number of papers using that feature type together with that primary representation (multi-label over rows). LLMs = LLM-based encoders; BERT/ST = BERT or Sentence-transformer; Classic = classic distributional embeddings; Multi = multimodal representations; SeqUser = sequential user representations; Hybrid = hybrid content–collaborative; Tabular = tabular/attribute-based; Struct = structured reasoning representations.

Semantic Feature	LLMs	BERT/ST	Classic	Multi	Graph	SeqUser	Hybrid	Tabular	Struct
User-item interaction	10	2	3	0	3	2	4	1	2
Textual content	4	6	5	1	1	1	5	3	0
Similarity/Retrieval	5	7	4	6	2	1	2	0	2
Behavioral preference	0	4	0	1	0	0	0	2	0
Context/Interaction	5	2	1	1	1	2	0	0	2
Entity/Identifier	3	3	1	1	2	0	0	2	0
Fact–response alignment	3	4	0	0	0	0	0	2	1
Session behavior	0	1	0	0	0	4	1	1	0
Item metadata	2	1	0	0	0	0	1	0	0

5.5 Domain–Strategy Co-Occurrence (RQ5)

RQ5 examines how algorithmic strategies distribute across application domains. Table 8 summarizes the co-occurrence of domains and strategy families, revealing clear specialization patterns rather than a uniform use of techniques. In the *LLM Recommendation* domain, the dominant strategy is LLMs-based modeling: 27 papers treat a pretrained LLM as the main decision-making component, often with limited architectural modification. Only a small number of works in this domain adopt RAG-centric hybrid models (4 papers) or reinforcement learning and sequential reasoning (3 papers), and there is almost no use of representation alignment or structure-aware strategies. This concentration suggests that most LLM-focused recommendation work is still exploring relatively generic LLM-centric designs before systematically incorporating more specialized mechanisms. The *Recommendation system* domain shows a more heterogeneous strategy mix. Alongside LLMs-based modeling, this domain accounts for the majority of cold-start/personalization strategies (8 papers), as well as non-trivial use of representation alignment (4 papers), retrieval optimization (2), and graph-and structure-aware modeling (1). These systems often extend or retrofit classical recommenders with RAG components or LLM-based modules, reflecting a stronger emphasis on addressing data sparsity and personalization challenges in established RS settings.

Table 8: Co-occurrence of domains and strategy families. Each cell reports the number of papers in that domain using the corresponding strategy family. LLMs-based = LLM-centric modeling; RetrOpt = retrieval optimization; RL/Seq = reinforcement learning and sequential reasoning; Align = representation alignment and matching; RAG-hybrid = RAG-centric hybrid modeling; MM/CD = multi-modal and cross-domain extension; ColdStart = cold-start/personalization; Robust = robustness, debiasing and safety; GraphStr = graph-and structure-aware modeling.

Domain	LLMs	RetrOpt	RL/ Seq	Align	RAG-hybrid	MM/ CD	ColdStart	Robust	GraphStr
LLM recommendation	27	1	3	0	4	0	0	1	1
Recommendation system	4	2	4	4	2	2	8	2	1
Information retrieval	6	11	1	3	1	1	0	0	0
LLM prompt	8	3	0	2	0	1	0	1	0
Question answering	3	2	0	0	0	1	0	0	0
Cold start	2	1	3	0	0	0	3	0	0
Cross modal	0	0	0	3	0	2	0	0	1
Multi-modal	0	1	0	0	0	2	0	0	0
LLM hallucination	1	1	0	0	0	0	0	0	0
Other/unspecified	2	0	0	0	0	0	0	0	0
Privacy security	1	0	0	0	0	0	0	1	0
TSFMs	0	0	0	0	0	0	1	0	0
Cross-context backdoor	1	0	0	0	0	0	0	0	0

Information retrieval papers, by contrast, are heavily skewed towards retrieval optimization: 11 IR studies fall into this strategy family, and several others employ alignment and matching techniques (3 papers) or LLMs-based modeling (6). This confirms that IR remains the primary testbed for improving retrieval quality and ranking policies, and that many retrieval-oriented innovations have not yet been fully transferred to recommendation-specific domains. Other domains concentrate different subsets of strategies. LLM prompt work combines LLMs-based modeling with retrieval optimization and robustness considerations, but rarely explores RL or structural modeling. Multi-modal and cross-modal domains emphasize multi-modal/cross-domain extension and alignment strategies, consistent with their focus on fusing or translating

between modalities. Cold-start papers combine LLMs-based, RL/seq, and RAG-hybrid strategies but remain few in number, highlighting that explicit cold-start treatment in RAG-based recommender systems is still relatively rare. Security-and robustness-oriented domains (Privacy security, Cross-Context Backdoor Attacks) naturally align with robustness and LLMs-based strategies but contribute only a handful of studies. Looking across Table 8, specialized strategies such as retrieval optimization, RL/sequence decision-making, structure-aware modelling, and explicit robustness mechanisms are unevenly explored across domains. Retrieval optimization is concentrated in information retrieval, cold-start strategies in classical recommendation, and multimodal/cross-domain extensions in cross-modal and multi-modal settings, while LLM Recommendation remains dominated by generic LLM-centric modelling. This uneven coverage points to substantial room for cross-fertilisation—for example, importing IR-style retrieval optimisation into LLM Recommendation, applying graph-and structure-aware strategies more broadly, or systematically evaluating robustness mechanisms in mainstream RS domains. These gaps provide concrete opportunities for future work to test and refine RAG strategies under the domain-specific constraints that matter most for recommendation.

5.6 Datasets Based on Task

The datasets adopted in the 138 surveyed RAG/LLM-based systems can be broadly categorized according to task orientation. Retrieval and ranking-oriented benchmarks [66,71,72] are most frequently employed to evaluate document relevance and retrieval quality. Recommendation datasets such as MovieLens, Amazon Reviews, and Yelp dominate in systems integrating LLMs with collaborative filtering or ranking objectives.

Conversational and dialogue-based datasets [5,130] are used in conversational recommender systems and persuasive dialogue settings. Knowledge-intensive and reasoning benchmarks [58,107,145] including HotpotQA, StrategyQA, are commonly adopted for evaluating hallucination mitigation and multi-hop reasoning performance. In addition, several studies utilize domain-specific or synthetic datasets [56,143] to assess application-driven systems.

5.7 Metrics Based on Task

As shown in Table 9, **discounted ranking metrics** (e.g., NDCG@k) [11,12,130] constitute the largest proportion of evaluation strategies, accounting for 30 studies. Discounted ranking metrics evaluate retrieval performance by incorporating the positional importance of relevant items within a ranked list. Higher-ranked items are assigned greater weight, reflecting the practical assumption that users are more likely to interact with early results. These metrics are therefore position-sensitive and widely adopted in recommendation and information retrieval benchmarks. Representative measures include DCG (Discounted Cumulative Gain), which accumulates graded relevance scores discounted by rank position, and NDCG@K (Normalized DCG@K), which normalizes DCG by the ideal ranking to enable cross-query comparability. Variants such as nDCG@5, nDCG@10, nDCG@20, nDCG@50, and nDCG@100 specify evaluation cutoffs. These metrics reflect ranking quality by accounting for both relevance and position, aligning evaluation with practical search and recommendation settings. In contrast to position-aware measures, **Hit-based metrics** [16,22] assess retrieval performance under a binary relevance assumption, examining only whether at least one ground-truth item appears within the top-K ranked results. These metrics ignore positional differences inside the cutoff and focus exclusively on the existence of a correct item in the returned list. Representative forms include HR@K (Hit Rate@K), also denoted as Hit@K or HitRatio@K, which measure the proportion of queries for which at least one relevant item is contained in the top-K results. Fixed-cutoff variants such

as Hits@1, Hits@3, and Hits@10 correspond to specific values of K. HitRatio@1 evaluates whether the top-ranked item is relevant. These metrics provide a coarse yet intuitive indicator of top-K retrieval success by verifying the presence of a correct answer within the retrieved set.

Table 9: Evaluation metric families adopted in the 138 RAG/LLM-based systems, with corresponding paper identifiers.

Metric Family	Paper IDs	Number of Papers (N = 138)	Share of All Papers (%)
Hit-based metrics	[7,16,37,40–42,65,81,89–91,94,98,113,129,137]	17	12.3
Recall-based metrics	[6,23,44,45,49,53,59,82,87,99,105,111,123,132,140,144,147]	17	12.3
Discounted ranking metrics	[10–12,14,15,20,21,25,27,38,39,63,75,80,83,84,92,95,96,101,102,108,110,112,119,121,122,126,127,133]	30	21.7
Average precision	[26,57,66,70,79,125,128,134–137]	11	8.0
Mean reciprocal rank	[24,48,62,69,71,72,74,103,107,116,141]	11	8.0
Rating prediction/Regression	[13,18,46,55,60,86,114,115]	8	5.8
Classification metrics	[3,19,23,35,47,50,51,54,61,64,67,85,117,120,124,131,139,143,149]	19	13.8
QA/Structured task metrics	[4,9,68,77,106,130,138,142,145,146]	10	7.2
Text generation metrics	[5,58,73,100,145]	5	3.6
Diversity/Novelty	[8,109]	2	1.4
Fairness/Bias/Causal	[36,118]	2	1.4
Efficiency/System metrics	[56]	1	0.7
Correlation/Agreement	[76,93]	2	1.4
Domain specific mertic	[17,52,88]	3	2.2

Recall-based metrics [6,45,147] quantify the extent to which a system retrieves the complete set of ground-truth items within a predefined cutoff. Typical forms include Recall@K, also written as R@K, which calculates the fraction of true target items appearing within the top-K results. Cutoff-specific variants such as Recall@1, Recall@5, Recall@10, Recall@20, Recall@50, and Recall@100 specify different evaluation depths. Recall@sum aggregates recall values across multiple positions or sessions. Variants such as Low-degree Recall focus on items with limited interaction frequency, while Sparse Recall evaluates performance under data sparsity conditions. These measures emphasize retrieval completeness by assessing how thoroughly the candidate list captures the ground-truth item set within the examined range. **Average precision-based metrics** [66,134] evaluate ranking quality by examining precision values at successive positions where relevant items occur. Instead of focusing solely on top-K inclusion or cumulative gain, these measures summarize the trade-off between precision and recall across the ranked list, rewarding systems that retrieve relevant items early and consistently. Core measures include AP (Average Precision), which computes the mean of precision scores at each rank containing a relevant item. MAP (Mean Average Precision) or mAP represents the average AP across all queries, enabling system-level comparison. MAP@K restricts the computation to the top-K results. This family captures ranking effectiveness by aggregating precision behavior across positions,

emphasizing early accurate retrieval while reflecting overall precision-recall balance. Closely related yet more position-focused, **Mean Reciprocal Rank (MRR)-based metrics** [24,48,62] measure ranking performance by focusing on the position of the first correct item in the returned list. The reciprocal of the rank at which the earliest relevant item appears is computed for each query, thereby assigning substantially higher scores to systems that place a correct result near the top. Sensitivity is therefore concentrated on the earliest successful match rather than on overall coverage or cumulative gain. Representative variants include MRR, as well as cutoff-restricted forms such as MRR@5, MRR@10, MRR@20, and MRR@40, which limit evaluation to a predefined ranking depth. Extensions such as Low-degree MRR examine performance on infrequent or long-tail items, while Sparse MRR evaluates robustness under sparse interaction settings. A total of 11 studies are categorized under this metric family. Shifting from ranking to score estimation, **Rating prediction and regression metrics** [13,60,114] assess the accuracy of continuous score estimation or probabilistic outcome prediction. Error-based measures include MSE (Mean Squared Error), which computes the average squared difference between predicted and true values; RMSE (Root Mean Squared Error), defined as the square root of MSE for scale interpretability; and MAE (Mean Absolute Error), which averages absolute deviations. LogLoss evaluates probabilistic predictions by penalizing confident but incorrect classifications. Ranking-oriented probabilistic indicators include AUC (Area Under the ROC Curve). Eight studies fall under this metric category. **Classification metrics** quantify predictive correctness in discrete label assignment tasks. Evaluation focuses on agreement between predicted categories and ground-truth labels, either through overall correctness, class-specific discrimination, or harmonic aggregation of precision and sensitivity. Accuracy measures the proportion of correctly predicted instances among all samples. Precision reflects the fraction of predicted positives that are truly positive, whereas Recall captures the proportion of actual positives successfully identified. F1-score represents the harmonic mean of precision and recall, balancing false positives and false negatives. Micro-F1 aggregates contributions across all instances prior to computing the F1 value, favoring frequent classes, while Macro-F1 averages class-wise F1 scores to treat each category equally. KF1 extends F1-based evaluation to knowledge-aware or task-specific settings. Pairs-F1 evaluates pairwise prediction consistency. This category includes 19 studies. **QA and structured task metrics** [106,138] evalute output correctness by requiring exact correspondence between predicted responses and ground-truth answers or executable results. Evaluation emphasizes strict matching rather than partial overlap or graded similarity. Exact Match (EM) measures the proportion of predictions that precisely match the reference answer at the token or string level, commonly applied in extractive question answering. Execution Accuracy (EX) evaluates structured outputs such as logical forms, SQL queries, or programs by verifying whether execution yields the correct result. This metric family captures deterministic correctness in answer generation and structured reasoning tasks. **Text generation metrics** [5,58] evaluate the quality of generated text by measuring overlap between model outputs and reference sentences. Emphasis is placed on n-gram matching to approximate fluency and content fidelity. BLEU (Bilingual Evaluation Understudy) computes the geometric mean of modified n-gram precision scores with a brevity penalty. BLEU-1 considers unigram overlap, while BLEU-2 extends evaluation to bigram consistency. This category includes 5 studies.

Diversity and novelty metrics assess the breadth and exploratory characteristics of recommendation outputs beyond accuracy-oriented evaluation. Attention is directed toward distributional balance, exposure of underrepresented items, and variation within or across recommendation lists. This category includes 2 studies. **Fairness and bias metrics** evaluate whether model outputs reflect equitable treatment across groups and whether observed effects remain robust under confounding factors. Emphasis is placed on counterfactual reasoning and distributional correction rather than pure predictive accuracy. Policy value estimates the expected utility of a learned decision policy under logged or simulated interaction data, often grounded in off-policy evaluation. Bias analysis (IPW adjustment) applies inverse propensity weighting to

correct exposure or selection bias, enabling unbiased performance estimation under observational data. This category includes 2 studies.

Efficiency and system metrics quantify computational cost and operational performance during model deployment or inference. Attention is directed toward responsiveness, resource consumption, and scalability under practical workloads. Latency, Query Time, Processing Time, and Running Throughput reflects the number of requests processed per unit time. GPU utilization captures hardware resource usage during execution. Speedup ratio compares performance gains relative to a baseline system. Memory usage assesses runtime memory consumption. API cost and Evaluation cost estimate monetary or computational expenditure associated with model operation. This category includes 1 study. **Correlation and agreement metrics** evaluate statistical consistency between predicted outcomes and reference signals, or between multiple evaluators or ranking outputs. Emphasis is placed on relational alignment rather than absolute accuracy. Pearson correlation measures linear association between continuous variables. Spearman correlation assesses monotonic rank-order relationships. Ranking consistency evaluates the stability or concordance of item ordering across models or conditions. This category includes 2 studies.

Domain-specific metrics capture evaluation criteria tailored to the unique objectives and constraints of a particular application domain. Unlike generic accuracy or ranking measures, these metrics operationalize task-specific goals, embedding domain knowledge directly into the evaluation protocol. For instance, study [88] specialize indicators such as ingredient-level sustainability scores and recipe-level sustainability functions are introduced to quantify environmental impact beyond conventional relevance metrics. Similarly, in reinforcement learning-based recommender systems, long-term return and policy value serve as domain-aligned objectives that reflect cumulative user engagement rather than immediate accuracy [17]. In retrieval-augmented generation settings, cost-aware objectives such as compression ratio-constrained likelihood optimization formalize efficiency-quality trade-offs specific to RAG pipelines [52]. Such metrics are defined by the structural properties, optimization targets, and practical constraints of the target system, ensuring that evaluation remains aligned with real-world deployment goals rather than generic benchmark performance.

6 Discussion

This section synthesizes the main findings of our systematic review and reflects on their implications for retrieval-augmented, LLM-based recommender systems. We structure the discussion around three themes that cut across RQ1–RQ5: (i) concentration of research in a few domains and strategy families, (ii) emerging but uneven use of semantic features and representations, and (iii) methodological gaps in evaluation and comparative analysis.

The domain-level analysis (RQ1) shows that most studies operate in familiar settings such as general recommendation and information retrieval, with LLM Recommendation and conventional Recommendation system accounting for the majority of the corpus, and Information retrieval contributing a substantial supporting role. More specialized domains—cold-start recommendation, multi-modal and cross-modal tasks, hallucination mitigation, security, and time-series modeling—appear only in small numbers. The domain-strategy matrix (RQ5) further reveals that algorithmic strategies are unevenly distributed: LLMs-based modeling dominates in LLM Recommendation, retrieval optimization is concentrated in IR, cold-start strategies appear mostly in classical recommendation, and multi-modal/cross-domain extensions cluster in cross-modal applications. These patterns suggest that current work prioritizes validating LLM-based techniques where data and benchmarks are abundant, while exploration of more complex or risk-sensitive domains is still limited. From the perspective of semantic techniques (RQ2) and feature-representation co-occurrence (RQ4), the review points to a hybrid but imbalanced semantic landscape. Interaction and

behavioral signals, textual content, similarity/retrieval features, contextual cues, and item metadata are all used, yet their prevalence and representation choices differ markedly. Interaction and session-level signals are typically paired with expressive encoders—LLMs, sequential models, or graph-based representations—reflecting their importance for fine-grained personalization under sparsity. Textual and similarity features are spread across both heavyweight (LLM, BERT/Sentence-transformer) and lightweight (classic embeddings, tabular encodings) pipelines, consistent with longstanding retrieval practice. At the same time, several combinations remain underexplored: few systems systematically investigate how alternative representations affect the usefulness of specific semantic signals, and lightweight or domain-specialized encoders play a relatively minor role compared with monolithic LLM backbones. This imbalance indicates that feature and representation design is still guided more by convenience and available models than by principled co-design for recommendation.

The algorithmic strategies reviewed in RQ3 also show a strong skew toward generic LLM-centric designs. Most systems rely on pretrained LLMs for preference elicitation, reasoning, or candidate refinement, often with minimal architectural changes. Retrieval optimization, RAG-centric hybrid modeling, reinforcement learning and sequential decision-making, graph-and structure-aware modeling, representation alignment, multi-modal/cross-domain extensions, and explicit robustness mechanisms appear, but each in a smaller subset of studies. When combined with the domain-strategy analysis (RQ5), this suggests that many promising ideas remain siloed: retrieval optimization and alignment techniques mature in IR; graph-based models and cold-start strategies develop largely within classical recommendation; and robustness-oriented methods cluster around security or hallucination-focused work. Systematic transfer of these strategies into LLM-based recommendation remains limited. Several methodological gaps cut across all five research questions. First, only a minority of studies provide detailed documentation of dataset properties such as user-item sparsity, long-tail distributions, and domain-specific biases, which makes it difficult to judge how well the reported methods would generalize beyond the chosen benchmarks. Second, evaluation protocols are often heterogeneous: works differ in their choice of metrics, baselines, and experimental settings, and few papers explicitly evaluate performance on tail subsets, cold-start conditions, or out-of-domain scenarios. Third, very few studies compare multiple algorithmic strategies or feature–representation combinations under controlled conditions. As a result, the field currently lacks robust evidence about which design choices are most effective for long-tail recommendation, hallucination mitigation, or dynamic user alignment.

These observations highlight both progress and open opportunities. On the positive side, there is clear movement toward semantically grounded recommendation: systems increasingly integrate rich behavioural, textual, and contextual features with retrieval-augmented LLMs, and early work on hybrid, structure-aware, and robustness-oriented strategies provides a foundation for more sophisticated pipelines. At the same time, the concentration of research in a small number of domains, the reliance on generic LLM-centric representations, and the scarcity of systematic evaluations under sparse or shifting conditions indicate that the design space is far from exhausted. Future work on RAG-based recommender systems would benefit from (i) more deliberate co-design of semantic features and representations, (ii) broader deployment of retrieval optimization, structure-aware modeling, and robustness mechanisms in mainstream recommendation settings, and (iii) standardized, transparency-oriented evaluation protocols that make long-tail, cold-start, and domain-shift performance first-class evaluation targets. This review intentionally focuses on venues from 2023 to 2025 due to the exceptionally rapid evolution of retrieval-augmented LLM research. The architectural and semantic design space of RAG-based recommender systems has undergone substantial transformation within a short time frame. By restricting the scope to recent high-impact publications, we aim to capture the most current methodological advances and emerging design patterns. While this may exclude certain

earlier foundational or industrial reports, the decision reflects the field's fast-paced progression rather than an oversight of prior work.

6.1 2025 Industrial Implications RAG-Based Recommender Systems

The RAG-IoE framework [150] proposes a retrieval-augmented generation architecture tailored for industry 5.0 environments within the internet of everything paradigm. The approach addresses limitations of conventional RAG pipelines in industrial contexts where heterogeneous enterprise data, contextual constraints, access control requirements, operational safety considerations coexist. A central design principle concerns context derivation from dynamic activity states rather than static user queries. Task descriptions, operational conditions, spatial positioning serve as implicit query sources. Retrieval objectives thus shift from general relevance toward situational compatibility. This design reduces irrelevant candidate exposure while preserving semantic flexibility. Deployment considerations emphasize hybrid infrastructure. Structured enterprise knowledge remains centrally maintained. Generative inference may operate in distributed or edge environments to mitigate latency, privacy exposure, infrastructure dependency. Such configuration supports scalability under industrial data governance constraints. Evaluation combines structured filtering accuracy with semantic retrieval effectiveness, contextual relevance assessment, expert validation procedures. Performance analysis demonstrates feasibility of hybrid symbolic–semantic retrieval under realistic industrial workflows. The RAG-IoE architecture reframes retrieval-augmented generation as a knowledge orchestration mechanism for controlled decision support in enterprise systems. Emphasis is placed on governance-aware filtering, contextual adaptability, infrastructure feasibility, operational robustness.

An adaptive RAG-based question-answering system in the context of industry 5.0 [151] proposes an adaptive retrieval-augmented generation architecture tailored to human-centric industrial environments. The framework addresses limitations of conventional static RAG pipelines under dynamic manufacturing contexts characterized by heterogeneous enterprise data, operational constraints, cost sensitivity, multilingual requirements. The proposed architecture incorporates a routing mechanism that classifies incoming queries to determine whether internal vector-store retrieval or external web-based search is required. This adaptive selection mechanism enhances robustness in scenarios where enterprise knowledge bases lack sufficient coverage. The system integrates FAISS-based dense retrieval, Mistral 7B language modeling, LangGraph workflow orchestration, open-source embedding models, forming a cost-efficient infrastructure suitable for Industry 5.0 deployment. A hybrid hallucination detection strategy is introduced to improve factual consistency. Binary classification evaluates alignment between generated responses and retrieved evidence. Multi-model LLM-based evaluation further assesses output reliability using a structured agreement scale. Correlation analysis indicates strong consistency between automated evaluation outputs and human assessment, suggesting improved trustworthiness under industrial QA conditions.

The study [152] emphasizes multilingual capability, resource efficiency, governance-aware deployment feasibility. The adaptive RAG design is positioned as an alternative to computationally expensive fine-tuned models, offering scalability under enterprise constraints. The framework reframes RAG-based question answering as an operational decision-support mechanism aligned with industry 5.0 principles of human-centric intelligence, sustainability, resilience. This study introduces a retrieval-augmented generation (RAG)-enhanced generative AI chatbot designed to facilitate interactive decision-making within industry 5.0 environments. Industry 5.0 emphasizes human-centric manufacturing, sustainability, resilience, personalized production. Increasing industrial data complexity creates cognitive burdens for human operators, necessitating context-aware AI systems capable of integrating heterogeneous knowledge sources. The proposed framework integrates policy documents, academic literature, industry reports, real-time web data into a unified retrieval architecture. Semantic vector search is combined with external search APIs

to capture both static domain knowledge and dynamic updates. Retrieved content is re-ranked prior to prompt construction for a large language model, enhancing contextual accuracy, interpretability, response relevance. System implementation employs modular orchestration platforms, vector embedding models, high-performance vector databases. Text preprocessing includes semantic chunking with controlled overlap to preserve contextual continuity. Query vectors are matched via similarity search within the vector store. Retrieved snippets are parsed, merged with user input, transformed into structured prompts guiding controlled generation. Evaluation adopts a multi-dimensional assessment framework covering efficiency, accuracy, relevance, user satisfaction.

Real-Time RAG [153] for the identification of supply chain vulnerabilities presents a retrieval-augmented framework designed to minimize latency between real-world supply chain events and large language model response capability. Conventional LLMs remain constrained by static pretraining cutoffs, limiting responsiveness to emerging disruptions. The study formalizes a timeliness-quality optimization objective that minimizes the temporal gap between event occurrence and model-generated analysis under a defined performance threshold. The proposed architecture integrates automated web-scraping of regulatory disclosures, particularly SEC EDGAR filings, with a modular RAG pipeline. A dynamic ingestion mechanism continuously updates the vector store, enabling near-real-time incorporation of newly published documents. The objective function models joint optimization of retrieval probability and generation probability while maintaining output quality above a target metric threshold. Empirical evaluation employs ranking metrics including nDCG, Hit Rate, Average Rank, MRR for retrieval quality, ROUGE, BLEU, Exact Match, Semantic Similarity for generation quality. Trade-space analysis demonstrates that fine-tuning the retriever yields the largest performance gains relative to latency cost. Adaptive iterative retrieval further enhances effectiveness under complex queries. Generator fine-tuning produces marginal improvements relative to computational overhead. The results underscore the central role of robust retrieval in real-time industrial intelligence environments. RAG-oriented architectures are framed as scalable infrastructures capable of supporting automated regulatory assessment across domains such as national security, economic resilience, and supply chain governance. Within these deployment contexts, refinement of the retrieval stage functions as the primary mechanism for reconciling rapid response requirements with analytical reliability.

6.2 Architectural Distinctions between RAG and GraphRAG

Conventional RAG architectures combine a parametric language model with dense vector retrieval over unstructured document chunks, where relevance is determined through similarity in embedding space [154,155]. Retrieved passages are treated as independent evidence units, leaving relational dependencies implicitly encoded within representations rather than explicitly modeled. This flat retrieval paradigm prioritizes scalability and efficient approximate nearest neighbor search, proving effective for open-domain question answering and document-grounded generation [156].

GraphRAG architectures introduce an explicit structural layer, organizing entities, concepts, or document segments into graph topologies prior to retrieval [157,158]. Retrieval shifts from similarity-based ranking toward subgraph extraction or neighborhood expansion, enabling multi-hop aggregation over relational paths. Structural reasoning becomes traceable through graph traversal, enhancing compositional inference and contextual coherence. Increased modeling capacity, however, incurs additional graph construction and traversal overhead.

Architectural divergence thus reflects a distinction between similarity-driven flat retrieval and topology-aware relational retrieval. The former optimizes scalability within large unstructured corpora; the latter strengthens reasoning depth in structured or knowledge-intensive settings.

6.3 Consistency of Metrics

Although a broad spectrum of evaluation metrics is observed across the surveyed studies, the overall metric ecosystem does not exhibit strong structural consistency. A clear concentration on relevance-oriented ranking metrics—particularly NDCG, MAP, MRR, and HR—indicates that most RAG/LLM-based systems are assessed primarily through the lens of top-K retrieval effectiveness. This dominance suggests that evaluation practices remain heavily influenced by traditional information retrieval and recommender system paradigms. However, the objectives of RAG-based systems extend beyond ranking accuracy. Such systems inherently combine retrieval, reasoning, and generation components, yet generation fidelity and faithfulness are comparatively underrepresented in the evaluation landscape. Text generation metrics appear in only a small subset of studies, and robustness-or hallucination-oriented indicators are rarely standard-ized. This reveals a partial misalignment between system architecture and evaluation focus. Moreover, metric redundancy is noticeable. Under single ground-truth settings, HR@K and Recall@K often converge numerically, while MRR@1 approximates Hit@1. Similarly, MAP and NDCG partially overlap in rewarding early correct placements. Despite this functional similarity, multiple closely related metrics are frequently reported simultaneously, which may inflate perceived evaluation diversity without substantially expanding analytical perspective. An additional imbalance emerges in the limited adoption of diversity, fairness, and efficiency metrics. While ranking accuracy dominates evaluation practice, system-level considerations such as latency, computational cost, and deployment scalability remain marginal. Collectively, prevailing evaluation practices exhibit a pronounced emphasis on relevance-oriented performance, with comparatively weak integration across complementary assessment dimensions. This pattern indicates partial misalignment between commonly reported metrics and the holistic objectives of end-to-end RAG systems. A more unified evaluation paradigm calls for systematic cross-family normalization, transparent recognition of metric redundancy, and proportionate inclusion of robustness, efficiency, fairness, and broader societal considerations. Absent such structural coordination, cross-study comparisons risk remaining analytically disjointed, even when methodological variation appears substantial.

6.4 Quantitative Synthesis of Reported Performance Gains

Although the surveyed literature frequently reports performance improvements over baselines, direct cross-paper effect-size meta-analysis remains difficult due to heterogeneity in datasets, task formulations, evaluation protocols, baseline choices, and reporting conventions. In particular, identical metric names often correspond to different cutoffs, candidate-set construction, or ground-truth definitions, which limits strict numerical comparability across studies.

Across retrieval and ranking-oriented evaluations, improvements most commonly appear in top-K relevance metrics such as Recall@K, NDCG@K, HR@K, and MRR@K. Reported gains frequently fall into low-to-moderate ranges in absolute terms, while relative improvements become larger in sparse, cold-start, or long-tail subsets. Several studies report consistent uplift across both Recall and NDCG, indicating that retrieval augmentation and re-ranking interventions tend to improve candidate coverage while also improving early-rank quality. In contrast, generation-or reasoning-oriented tasks typically report gains in EM, EX, or F1, often alongside reductions in hallucination-related error indicators when such measurements are provided. Regression-and prediction-driven systems mainly report improvements in RMSE/MAE or AUC-family metrics, with gains often presented as consistent but modest increases over strong supervise d baselines.

A comparative pattern emerges at the strategy level. Retrieval-side interventions (e.g., improved indexing, hard-negative training, better retriever training, candidate expansion) most reliably translate into improvements in ranking metrics, especially under large-corpus settings. Generation-side interventions

(e.g., alignment, reranking with LLM feedback, constrained decoding, multi-step prompting) more strongly affect answer correctness or faithfulness indicators, though standardized reporting remains limited. Hybrid strategies combining retrieval tuning with generation control frequently show the most stable improvements across multiple metric families, suggesting that end-to-end gains depend on coordinated optimization across pipeline stages rather than isolated component upgrades.

6.5 Cold-Start and Long-Tail

Cold-start and long-tail challenges constitute central research themes in recommender systems, yet explicit and standardized evaluation under these conditions remains limited within retrieval-augmented LLM-based recommendation frameworks. A small number of recent studies directly address item cold-start or sparse-interaction settings within RAG architectures. For instance, RAGSys: Item-Cold-Start Recommender as RAG System [159] formulates item cold-start recommendation as a retrieval-augmented problem, emphasizing the role of high-quality retrieval demonstrations in mitigating sparse supervision. Similarly, study [160] introduce a knowledge graph retrieval-augmented generation framework with explicitly constructed MovieLens cold-start splits, demonstrating that structured retrieval enhances recommendation robustness under unseen-item scenarios. Reference [2] suggest CoRAL, a collaborative retrieval-augmented LLM framework that explicitly targets long-tail recommendation by integrating user–item interaction evidence into the prompting process. Despite these targeted efforts, systematic long-tail stratification remains underdeveloped across the broader literature. While sparse or unseen conditions receive occasional experimental attention, consistent head-tail disaggregation, popularity-stratified reporting, and exposure-aware evaluation protocols are not widely adopted. Broader analyses of LLM-based recommender challenges [161] recognize sparse interaction and tail distribution phenomena as structural bottlenecks, yet empirical evaluation standards for RAG-based systems remain heterogeneous.

This imbalance suggests a methodological gap between architectural claims of improved generalization and the empirical conditions under which performance is validated. Aggregate improvements in Recall@K, NDCG@K, or HR@K do not necessarily indicate robustness in cold-start or long-tail regimes unless accompanied by explicit distribution-aware reporting. Retrieval augmentation shows potential for mitigating memorization gaps and enhancing unseen-item generalization; however, the absence of standardized cold-start simulation and long-tail exposure metrics limits conclusive interpretation of such gains. A more rigorous evaluation paradigm would require controlled unseen-entity splits, popularity-stratified analysis, and explicit tail-exposure measurement to align empirical validation with the structural objectives of retrieval-augmented recommender architectures.

6.6 User-Centric and Interactive Evaluation Limitations

A noticeable concentration on offline performance metrics characterizes the surveyed literature. Ranking-oriented measures, prediction accuracy, and benchmark-based evaluation dominate empirical validation practices. In contrast, user-centered assessment—including controlled user studies, explanation quality evaluation, trust calibration analysis, and interactive feedback modeling—remains comparatively limited. This imbalance suggests a structural gap between algorithmic benchmarking and real-world recommendation dynamics. Offline improvements in NDCG, Recall, or AUC do not necessarily translate into enhanced perceived relevance, trustworthiness, or long-term engagement. Moreover, retrieval-augmented systems introduce additional explainability and faithfulness challenges that are insufficiently captured by conventional metrics.

Future evaluation frameworks may require multi-dimensional integration, combining offline effectiveness, user-perceived utility, explanation alignment, trust stability, and adaptive feedback loops to better reflect real-world recommender system deployment.

6.7 System-Level Operational Considerations beyond Scope

Operational considerations such as inference overhead, retrieval latency, horizontal scalability, external knowledge base maintenance, and energy consumption constitute essential components of real-world recommender system deployment. These dimensions significantly affect industrial viability, operational stability, and long-term sustainability of retrieval-augmented architectures. The analytical focus of this review centers on architectural configurations, algorithmic mechanisms, dataset utilization, and evaluation practices reported in the academic literature. A comprehensive systems-engineering assessment encompassing infrastructure expenditure, runtime optimization, or lifecycle maintenance strategies extends beyond the intended scope of the present study.

Subsequent research may advance the field by incorporating unified system-level benchmarking paradigms that evaluate not only predictive effectiveness but also computational efficiency and sustainability, particularly within large-scale production environments.

6.8 Adoption Frequency vs. Practical Optimality

High frequency of adoption does not necessarily indicate architectural optimality. The prevalence of LLM-centric strategies in recent literature reflects rapid technological diffusion and research attention rather than universal efficiency or deployment suitability. LLM-augmented pipelines often incur increased computational cost, latency overhead, and infrastructure complexity compared to lightweight retrieval or embedding-based approaches. In large-scale production environments, such trade-offs may significantly affect feasibility.

Therefore, adoption frequency should be interpreted as a signal of research emphasis rather than definitive empirical validation. Balanced evaluation requires simultaneous consideration of effectiveness, efficiency, and scalability constraints.

7 Future Research Directions

Our findings point to several concrete directions for advancing retrieval-augmented, LLM-based recommender systems. First, there is a need for more deliberate design of retrieval policies and retrieval-aware architectures. While similarity and retrieval features are widely used, only a small subset of systems implement explicit retrieval optimization or policy learning, and these are concentrated in information-retrieval settings rather than recommendation. Future work could extend retrieval optimization, reinforcement learning, and other adaptive control mechanisms to mainstream recommendation tasks, with a focus on long-tail and cold-start conditions where retrieval choices are most critical. This includes learning when to retrieve, how many candidates to request, and how to balance log-based and external knowledge sources for different user states. Second, the co-occurrence patterns between semantic features and representation families suggest that representation learning remains heavily LLM-centric. Interaction and contextual features are commonly assigned to LLM encoders or sequential models, while textual and similarity features are spread across BERT-style and classic embeddings. Lightweight, graph-based, multimodal, and hybrid representations are present but underutilized. Future research should more systematically explore feature-representation co-design: for example, combining graph or tabular encodings with LLMs for cold-start users and items, or pairing session-level features with multimodal representations in multimedia recommendation. Controlled comparisons of alternative encoders for the same semantic feature types would help

clarify when heavy LLM backbones are necessary and when simpler models suffice. Third, several domains identified in our review remain underexplored, despite being highly relevant to practical deployment. Cold-start recommendation, cross-and multi-modal recommendation, hallucination mitigation, privacy and security, and time-series forecasting together account for only a small fraction of the corpus. Transferring mature strategies from better-studied domains into these settings—for example, applying IR-style retrieval optimization to cold-start recommendation, or graph-and structure-aware modeling to security-sensitive scenarios—offers a promising avenue. At the same time, new benchmarks and datasets that better reflect these challenging domains are needed to move beyond proof-of-concept evaluations.

Finally, methodological and evaluation practices require more attention. Many studies provide limited information about user-item sparsity, long-tail distributions, and domain-specific biases, and few explicitly evaluate performance on tail subsets, cold-start users/items, or out-of-domain scenarios. Standardized reporting of dataset characteristics, along with evaluation protocols that include long-tail and cold-start test splits by default, would make it easier to compare methods and assess robustness. In addition, cross-strategy comparisons—e.g., LLM-centric vs. RAG-hybrid vs. RL-based pipelines under the same data conditions—are largely missing from the current literature. Designing such comparative studies is essential for understanding which combinations of semantic features, representations, and algorithmic strategies are most effective in realistic, data-constrained environments. These directions point beyond simply plugging LLMs into existing recommenders. They call for retrieval-augmented systems that are designed around the interplay between semantic signals, representation choices, and algorithmic strategies, and that are evaluated under conditions that reflect the sparsity, long-tail structure, and domain shifts encountered in practice.

8 Research Limitations

Despite adhering to a structured review protocol, this study has several important limitations. First, the initial screening phase relies on title and abstract relevance, a necessary compromise given the volume of retrieved records. This approach may omit studies whose contributions align with retrieval-augmented recommendation but are not explicitly reflected in metadata, which poses a threat to recall. Second, to keep the review feasible and focused on recent advances, we limit coverage to a predefined set of peer-reviewed venues (seven major conferences and six major journals) and a three-year window (2023–2025). This scoping decision improves topical focus but may omit relevant work published in other outlets or outside the chosen time window, including foundational work that predates current LLM-centric pipelines. Third, the analysis is constrained by inconsistent reporting practices across primary studies. Many works omit key dataset descriptors (e.g., sparsity levels, long-tail ratios, popularity distributions) and provide limited details on evaluation configurations, which reduces the reliability of cross-paper comparisons. Finally, the rapid pace of LLM-related research makes any synthesis time-sensitive; therefore, our conclusions represent a snapshot of an evolving field rather than an exhaustive account.

9 Conclusion

This systematic literature review examined how retrieval-augmented architectures and large language models are being used to support recommendation, including both LLM-based and conventional recommender systems. Using a three-axis framework that distinguishes domain characteristics, semantic feature and representation choices, and algorithmic strategy families, we synthesized 138 recent studies from leading conferences and journals and answered five research questions on domains, semantic techniques, representations, and strategy distributions.

The review shows that current work is concentrated in general recommendation and information-retrieval settings, while domains such as cold-start recommendation, multimodal and cross-modal tasks,

robustness, and security remain comparatively underexplored. Similarity and retrieval signals, user-item interaction semantics, and textual content form the core evidence sources, and LLM and BERT/Sentence-Transformer encoders dominate the representation space, with graph-based, multimodal, and hybrid representations used more selectively. At the strategy level, most systems still rely on generic LLM-centric modelling, whereas retrieval optimisation, reinforcement learning, structure-aware modelling, and explicit robustness mechanisms appear in smaller, domain-specific clusters.

These patterns indicate that the field is moving toward semantically grounded, retrieval-aware recommendation, but that key parts of the design space are still underused. The taxonomy and co-occurrence analyses provided in this survey can serve as a reference for future work on retrieval-augmented recommender systems, and they point to concrete directions for progress, including better alignment between semantic features and representations, broader deployment of specialised strategies in mainstream recommendation domains, and more systematic evaluation under long-tail, cold-start, and domain-shift conditions.

Acknowledgement: The authors acknowledge that this work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00438056), and by the Chung-Ang University Research Scholarship Grants in 2024. This work also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02305436, Development of Digital Innovative Element Technologies for Rapid Prediction of Potential Complex Disasters and Continuous Disaster Prevention).

Funding Statement: This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00438056), and by the Chung-Ang University Research Scholarship Grants in 2024. This work also supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02305436, Development of Digital Innovative Element Technologies for Rapid Prediction of Potential Complex Disasters and Continuous Disaster Prevention).

Author Contributions: Minhyeok Choi and Imran Ahsan lead the conceptualization of the study. Minhyeok Choi developed the methodology, implemented the software, conducted the investigation, performed the formal analysis, curated the data, managed the resources, and prepared the original draft of the manuscript. Validation of the results is carried out jointly by Minhyeok Choi, Imran Ahsan, Hyunwook Yu, and Taeyoung Choe. Imran Ahsan additionally contributed to manuscript review, discussion of analytical results, and verification of methodological consistency. Minhyeok Choi is also responsible for the visualization and for the review and editing of the manuscript. Supervision and oversight of the research activities, including project administration and funding acquisition, are provided by Mucheol Kim. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmc.2026.079504/s1.

Appendix A Information Sources and Search Strategy

For each venue, we covered all available issues/proceedings within 2023–2025; some sources had limited availability for specific years, as shown in Table A1.

Appendix A.1 Information Sources, Coverage Years and Last-Accessed Dates

Table A1: Information sources, coverage years and official access points. We conducted record collection and screening between July and November 2025 and performed a final verification pass in December 2025 (last accessed dates shown).

Venue	Official Access Point Used	Year	Last Accessed
EMNLP (ACL Anthology)	https://aclanthology.org/venues/emnlp/	2023–2024	[25 December 2025]
ACL (ACL Anthology)	https://aclanthology.org/	2023–2025	[21 December 2025]
SIGIR (ACM DL/proceedings)	https://dl.acm.org/doi/proceedings/10.1145/3726302	2023–2025	[25 December 2025]
RecSys (ACM DL/proceedings)	https://dl.acm.org/doi/proceedings/10.1145/3640457	2023–2024	[26 December 2025]
WSDM (ACM DL/proceedings)	https://dl.acm.org/doi/proceedings/10.1145/3701551	2023–2025	[28 December 2025]
KDD (ACM DL/proceedings)	https://dl.acm.org/doi/proceedings/10.1145/3637528	2023–2025	[28 December 2025]
NeurIPS (proceedings site)	https://neurips.cc/	2023–2024	[27 December 2025]
ESWA (journal portal)	https://www.sciencedirect.com/journal/expert-systems-with-applications	2023–2025	[28 December 2025]
IPM (journal portal)	https://www.sciencedirect.com/journal/information-processing-and-management	2023–2025	[25 December 2025]
UMUAI (journal portal)	https://link.springer.com/journal/11257	2023–2024	[28 December 2025]
ACM TORS (journal portal)	https://dl.acm.org/journal/tors	2023–2025	[24 December 2025]
TKDE (IEEE Xplore)	https://dl.acm.org/journal/ieeecs_tkde	2023–2025	[26 December 2025]
JAIR (journal portal)	https://www.jair.org/index.php/jair/issue/archive	2023–2025	[27 December 2025]

Appendix A.2 Concrete Query Strings and Keyword Filters

We use a three-block concept formulation and implement it using the following concrete keyword sets.

Concept block 1 (LLM terms).

```
("large language model" OR LLM OR "foundation model")
```

Concept block 2 (recommendation terms).

```
(recommend* OR "recommender system" OR "recommendation system"
OR "conversational recommendation" OR "sequential recommendation" OR
ranking)
```

Concept block 3 (retrieval augmentation terms).

```
("retrieval-augmented generation" OR RAG OR "retrieval augmented"
OR "semantic retrieval" OR retriev* OR "dense retrieval")
```

Venue-portal adaptations.

- **If the portal supports only simple queries:** we run each concept block separately and then intersect results during screening (title/abstract stage).
- **If the portal supports field filters:** we apply queries to Title/Abstract fields and limit years to 2023–2025.

- **If the portal provides only venue-year listings (TOC/proceedings pages):** we compile all records for the venue-year range and apply the above keyword filters during title/abstract screening.

Appendix A.3 Record Counts by Source

Table A2: Record counts by source (identification and final inclusion).

Source	Records Retrieved	Records Included
EMNLP	4127	7
ACL	4994	5
SIGIR	1026	16
RecSys	105	16
WSDM	223	10
KDD	908	12
NeurIPS	4096	4
Journals (ESWA/IPM/UMUAI/TORS/TKDE/JAIR)	8076	68
Total	**23,555**	**138**

Appendix B Full-Text Exclusions with Reasons

Appendix B.1 Exclusion Reason Codes

We assign one primary reason code to each excluded full-text paper:

- **E1: Not LLM-based** (no LLM/foundation model component)
- **E2: Not recommendation-related** (outside recommendation scope)
- **E3: Not retrieval augmentation** (no semantic retrieval/no RAG/no external evidence retrieval)
- **E4: Not peer-reviewed research article/insufficient full text** (e.g., editorial/tutorial/extended abstract)
- **E5: No quantitative outcomes** (insufficient quantitative evaluation for extraction)
- **E6: Insufficient methodological detail for coding** (cannot reliably label feature/representation/strategy variables)
- **E7: Duplicate/overlap** (duplicate record; merged into another entry)

Appendix C Data Extraction Template and Codebook

Appendix C.1 Extraction Form

Table A3: Data extraction template and coding schema used for this review.

Field	Definition/Coding Rule
PaperID	Unique study identifier used consistently across Tables 3–8 (e.g., [2–146]).
Year, Venue	Publication year and venue within the review scope (2023–2025; EMNLP/ACL/NeurIPS/KDD/SIGIR/RecSys/WSDM; ESWA/IPM/UMUAI/TORS/TKDE/JAIR).
Task type	Recommendation task setting (e.g., sequential recommendation, conversational recommendation, candidate generation, reranking/ranking).

(Continued)

Table A3 (continued)

Field	Definition/Coding Rule
Retrieval component	Retrieved artifact(s) (items/documents/reviews/knowledge); retrieval model type (BM25, dense/dual-encoder, hybrid, reranker); retrieval depth (top-k if reported); index/evidence store type (if reported).
LLM component	LLM usage locus (ranking, generation, explanation, planning); adaptation type (prompting/in-context learning/fine-tuning if reported); model name/family if explicitly reported by the study.
Domain label(s)	One or more application domains (multi-label allowed): *LLM Recommendation, Recommendation system, Information retrieval, LLM prompt, Question answering, Cold start, Cross modal, Multi-modal, LLM hallucination, Privacy/security, Time-series forecasting models (TSFMs), Cross-context backdoor attacks, Multi-hop reasoning, Other/unspecified*.
Semantic feature type(s)	One or more semantic feature types (multi-label allowed): *Similarity/Retrieval, User–item interaction/User behavior, Textual content, Context/Interaction, Entity/Identifier, Fact–response alignment, Behavioral preference, Session behavior, Item metadata*.
Primary representation family	Exactly one primary representation family (dominant backbone): *LLMs; BERT/Sentence-transformer; Classic distributional embeddings; Multimodal representations; Graph-based representations; Sequential user representations; Hybrid content–collaborative; Tabular/attribute-based; Structured reasoning representations*. Secondary families (if any) are recorded in Notes.
Strategy family (multi-label)	One or more algorithmic strategy families: *LLMs-based; Retrieval optimization; Representation alignment & matching; Graph & structure-aware modeling; RAG-centric hybrid modeling; Reinforcement learning & sequential reasoning; Robustness, debiasing & safety; Multi-modal & cross-domain extension; Cold-start/Personalization*.
Datasets and metrics	Dataset(s), evaluation protocol, and quantitative metrics (e.g., NDCG/Recall/HR@K; retrieval metrics; quantitative user-study results if reported).
Evidence pointer	Location cues that justify coding (section/page/paragraph pointer; short note describing the supporting evidence).

Appendix C.2 Codebook

Codebook entry (Semantic feature type: Similarity/Retrieval).

We assign *Similarity/Retrieval* when the method explicitly uses similarity signals (lexical or embedding-based) or retrieval scores as a semantic feature for candidate selection, ranking, grounding, or evidence selection (e.g., dense retrieval similarity, cosine/dot-product similarity, reranker relevance). If the method only encodes text without an explicit retrieval/similarity mechanism, we assign *Textual content* instead.

Decision rule (Primary representation family).

We assign the *primary* representation family as the dominant backbone used to encode the main semantic signals in the model (e.g., LLM encoder vs. BERT/Sentence-transformer vs. graph encoder). If a study uses multiple encoders, we record secondary encoders in notes but keep exactly one primary family for Table 5 consistency.

References

1. Arslan M, Ghanem H, Munawar S, Cruz C. A survey on RAG with LLMs. Procedia Comput Sci. 2024;246:3781–90. doi:10.1016/j.procs.2024.09.178.
2. Wu J, Chang CC, Yu T, He Z, Wang J, Hou Y, et al. Coral: collaborative retrieval-augmented large language models improve long-tail recommendation. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024 Aug 25–29; Barcelona, Spain. p. 3391–401.
3. Kim G, Kim S, Jeon B, Park J, Kang J. Tree of clarifications: answering ambiguous questions with retrieval-augmented large language models. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 996–1009.
4. Malik V, Jagatap A, Puranik VS, Majumder A. PEARL: preference extraction with exemplar augmentation and retrieval with LLM agents. In: Dernoncourt F, Preoțiuc-Pietro D, Shimorina A, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track; 2024 Nov 12–16; Miami, FL, USA. p. 1536–47.
5. Furumai K, Legaspi R, Romero JCV, Yamazaki Y, Nishimura Y, Semnani S, et al. Zero-shot persuasive chatbots with LLM-generated strategies and information retrieval. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov 12–16; Miami, FL, USA. p. 11224–49. doi:10.18653/v1/2024.findings-emnlp.656.
6. Huang F, Bei Y, Yang Z, Jiang J, Chen H, Shen Q, et al. Large language model simulator for cold-start recommendation. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25; 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 261–70. doi:10.1145/3701551.3703546.
7. Azizi A, Momtazi S. SNRBERT: session-based news recommender using BERT. User Model User-Adap Interact. 2024;34(4):1071–85. doi:10.1007/s11257-024-09409-x.
8. Hazrati N, Ricci F. Choice models and recommender systems effects on users' choices. User Model User-Adap Interact. 2024;34(1):109–45. doi:10.1007/s11257-023-09366-x.
9. Cremaschi M, Ditolve D, Curcio C, Panzeri A, Spoto A, Maurino A. Decoding the mind: a RAG-LLM on ICD-11 for decision support in psychology. Expert Syst Appl. 2025;279:127191. doi:10.1016/j.eswa.2025.127191.
10. Cui Y, Wang K, Yu H, Guo X, Cao H. KLLMs4Rec: knowledge graph-enhanced LLMs sentiment extraction for personalized recommendations. Expert Syst Appl. 2025;282(2):127430. doi:10.1016/j.eswa.2025.127430.
11. Luo S, Xu J, Zhang X, Wang L, Liu S, Hou H, et al. RALLRec+: retrieval augmented large language model recommendation with reasoning. Expert Syst Appl. 2025;297:129508. doi:10.1016/j.eswa.2025.129508.
12. Wei C, Duan K, Zhuo S, Wang H, Huang S, Liu J. Enhanced recommendation systems with retrieval-augmented large language model. J Artif Intell Res. 2025;82:1147–73. doi:10.1613/jair.1.17809.
13. Bai Z, Zheng Y, Yang P, Liu S, Zhang Y, Chang Y. DCRLRec: dual-domain contrastive reinforcement large language model for recommendation. Inform Process Manag. 2025;62(4):104140. doi:10.1016/j.ipm.2025.104140.
14. Boz A, Zorgdrager W, Kotti Z, Harte J, Louridas P, Karakoidas V, et al. Improving sequential recommendations with LLMs. ACM Trans Recomm Syst. 2026;4(2):1–35. doi:10.1145/3711667.
15. Tian C, Hu B, Gan C, Chen H, Zhang Z, Yu L, et al. ReLand: integrating large language models' insights into industrial recommenders via a controllable reasoning pool. In: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24; 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 63–73. doi:10.1145/3640457.3688131.

16. Li Y, Zhai X, Alzantot M, Yu K, Vulić I, Korhonen A, et al. CALRec: contrastive alignment of generative LLMs for sequential recommendation. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 422–32. doi:10.1145/3640457.3688121.
17. Wang J, Karatzoglou A, Arapakis I, Jose JM. Large language model driven policy exploration for recommender systems. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25; 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 107–16. doi:10.1145/3701551.3703496.
18. Yu X, Zhang L, Chen C. Explainable CTR prediction via LLM reasoning. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25; 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 707–16. doi:10.1145/3701551.3703551.
19. Heo S, Son S, Park H. HaluCheck: explainable and verifiable automation for detecting hallucinations in LLM responses. Expert Syst Appl. 2025;272:126712. doi:10.1016/j.eswa.2025.126712.
20. Song J, Wang X, Zhu J, Wu Y, Cheng X, Zhong R, et al. RAG-HAT: a hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In: Dernoncourt F, Preoţiuc-Pietro D, Shimorina A, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track; 2024 Nov 12–16; Miami, FL, USA. p. 1548–58. doi:10.18653/v1/2024.emnlp-industry.113.
21. Kemper S, Cui J, Dicarlantonio K, Lin K, Tang D, Korikov A, et al. Retrieval-augmented conversational recommendation with prompt-based semi-structured natural language state tracking. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24; 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 2786–90. doi:10.1145/3626772.3657670.
22. Di Palma D. Retrieval-augmented recommender system: enhancing recommender systems with large language models. In: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23; 2023 Sep 18–22; Singapore. New York, NY, USA: Association for Computing Machinery; 2023. p. 1369–73. doi:10.1145/3604915.3608889.
23. Maes U, Michiels L, Smets A. GenUI(ne) CRS: UI elements and retrieval-augmented generation in conversational recommender systems with LLMs. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 1177–9. doi:10.1145/3640457.3691697.
24. Kang S, Jin B, Kweon W, Zhang Y, Lee D, Han J, et al. Improving scientific document retrieval with concept coverage-based query set generation. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25; 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 895–904. doi:10.1145/3701551.3703544.
25. Ding Y, Wang B, Cui X, Xu M. Popularity prediction with semantic retrieval for news recommendation. Expert Syst Appl. 2024;247(2):123308. doi:10.1016/j.eswa.2024.123308.
26. Li J, Wong WK, Jiang L, Jiang K, Fang X, Xie S, et al. Collaboratively semantic alignment and metric learning for cross-modal hashing. IEEE Trans Know Data Eng. 2025;37(5):2311–28. doi:10.1109/tkde.2025.3537704.
27. Xi Y, Liu W, Lin J, Weng M, Cai X, Zhu H, et al. Efficient and deployable knowledge infusion for open-world recommendations via large language models. ACM Trans Recomm Syst. 2025;4(1):1–36. doi:10.1145/3725894.
28. Fan W, Ding Y, Ning L, Wang S, Li H, Yin D, et al. A survey on RAG meeting LLMs: towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 6491–501. doi:10.1145/3637528.3671470.
29. Church KW, Sun J, Yue R, Vickers P, Saba W, Chandrasekar R. Emerging trends: a gentle introduction to RAG. Nat Lang Eng. 2024;30(4):870–81. doi:10.1017/s1351324924000044.
30. Zhao S, Yang Y, Wang Z, He Z, Qiu LK, Qiu L. Retrieval augmented generation (RAG) and beyond: a comprehensive survey on how to make your LLMs use external data more wisely. arXiv:2409.14924. 2024.

31. Swacha J, Gracel M. Retrieval-augmented generation (RAG) chatbots for education: a survey of applications. Appl Sci. 2025;15(8):4234. doi:10.3390/app15084234.
32. Huang Y, Huang J. A survey on retrieval-augmented text generation for large language models. arXiv:2404.10981. 2024.
33. Li Z, Wang Z, Wang W, Hung K, Xie H, Wang FL. Retrieval-augmented generation for educational application: a systematic survey. Comput Educat Artif Intell. 2025;8:100417. doi:10.1016/j.caeai.2025.100417.
34. Hu Y, Lu Y. RAG and RAU: a survey on retrieval-augmented language model in natural language processing. arXiv:2404.19543. 2025.
35. Rajapakse D, Leith D. User cold-start learning in recommender systems using monte carlo tree search. ACM Trans Recomm Syst. 2025;3(1):1–23. doi:10.1145/3618002.
36. Silva N, Silva T, Werneck H, Rocha L, Pereira A. User cold-start problem in multi-armed bandits: when the first recommendations guide the user's experience. ACM Trans Recomm Syst. 2023;1(1):1–24. doi:10.1145/3554819.
37. Gong Z, Wu X, Chen L, Zheng Z, Wang S, Xu A, et al. Full index deep retrieval: end-to-end user and item structures for cold-start and long-tail item recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23; 2023 Sep 18–22; Singapore. New York, NY, USA: Association for Computing Machinery; 2023. p. 47–57. doi:10.1145/3604915.3608773.
38. Rajput S, Mehta N, Singh A, Hulikal Keshavan R, Vu T, Heldt L, et al. Recommender systems with generative retrieval. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates, Inc.; 2023. p. 10299–315.
39. Bao K, Zhang J, Zhang Y, Huo X, Chen C, Feng F. Decoding matters: addressing amplification bias and homogeneity issue in recommendations for large language models. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 10540–52. doi:10.18653/v1/2024.emnlp-main.589.
40. Yang S, Ma W, Sun P, Ai Q, Liu Y, Cai M, et al. Sequential recommendation with latent relations based on large language model. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24; 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 335–44. doi:10.1145/3626772.3657762.
41. Liu Q, Wu X, Wang Y, Zhang Z, Tian F, Zheng Y, et al. LLM-ESR: large language models enhancement for long-tailed sequential recommendation. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al., editors. Proceedings of the 38th International Conference on Neural Information Processing Systems; 2024 Dec 10–15; Vancouver, BC, Canada. Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 26701–27.
42. Xu X, Dong H, Qi L, Zhang X, Xiang H, Xia X, et al. CMCLRec: cross-modal contrastive learning for user cold-start sequential recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24; 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 1589–98. doi:10.1145/3626772.3657839.
43. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell System Rev. 2022;18(2):e1230. doi:10.1002/cl2.1230.
44. Ning LB, Wang S, Fan W, Li Q, Xu X, Chen H, et al. CheatAgent: attacking LLM-empowered recommender systems via LLM agent. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 2284–95. doi:10.1145/3637528.3671837.
45. Kim S, Kang H, Choi S, Kim D, Yang M, Park C. Large language models meet collaborative filtering: an efficient all-round LLM-based recommender system. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 1395–406. doi:10.1145/3637528.3671931.
46. Jeunen O, Goethals B. Pessimistic decision-making for recommender systems. ACM Trans Recomm Syst. 2023;1(1):1–27. doi:10.1145/3568029.

47. Alharbe N, Rakrouki MA, Aljohani A. A collaborative filtering recommendation algorithm based on embedding representation. Expert Syst Appl. 2023;215:119380. doi:10.1016/j.eswa.2022.119380.
48. de Campos LM, Fernández-Luna JM, Huete JF. Use of topical and temporal profiles and their hybridisation for content-based recommendation. arXiv:2401.10607. 2024.
49. Rendle S. Efficient optimization of sparse user encoder recommenders. ACM Trans Recomm Syst. 2024;2(3):1–31. doi:10.1145/3651170.
50. Otaki K, Baba Y. Travel itinerary recommendation using interaction-based augmented data. Expert Syst Appl. 2025;269(3):126294. doi:10.1016/j.eswa.2024.126294.
51. Wu D, Xiao E, Zhu Y, Jensen CS, Lu K. Efficient retrieval of the top-k most relevant event-partner pairs. IEEE Trans Know Data Eng. 2023;35(3):2529–43. doi:10.1109/tkde.2021.3118552.
52. Wu C, Shao N, Liu Z, Xiao S, Li C, Zhang C, et al. Lighter and better: towards flexible context adaptation for retrieval augmented generation. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25; 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 271–80. doi:10.1145/3701551.3703580.
53. Zou A, Hao W, Jin D, Zou S, Zheng Y, Sun F, et al. An efficient corpus indexer for dynamic corpora retrieval. Expert Syst Appl. 2024;254(3):124306. doi:10.1016/j.eswa.2024.124306.
54. Shao R, Tang Y, Yang L, Wang F. Law LLM unlearning via interfere prompt, review output and update parameter: new challenges, method and baseline. Expert Syst Appl. 2025;292(7):128612. doi:10.1016/j.eswa.2025.128612.
55. Li J, Liu W, Ding Z, Fan W, Li Y, Li Q. Large language models are in-context molecule learners. IEEE Trans Know Data Eng. 2025;37(7):4131–43. doi:10.1109/tkde.2025.3557697.
56. Sheng M, Wang S, Zhang Y, Wang K, Wang J, Luo Y, et al. MQRLD: a multimodal data retrieval platform with query-aware feature representation and learned index based on data lake. Inform Proces Manag. 2025;62(4):104101. doi:10.1016/j.ipm.2025.104101.
57. Liang M, Li Y, Yu Y, Cao X, Xue Z, Li A, et al. Structures aware fine-grained contrastive adversarial hashing for cross-media retrieval. IEEE Trans Know Data Eng. 2024;36(7):3514–28. doi:10.1109/tkde.2024.3356258.
58. He L, Tang P, Zhang Y, Zhou P, Su S. Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation. Inform Process Manag. 2025;62(4):104150. doi:10.1016/j.ipm.2025.104150.
59. Liu X, Wang R, Song Y, Kong L. GRAM: generative retrieval augmented matching of data schemas in the context of data security. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 5476–86. doi:10.1145/3637528.3671602.
60. Zhang H, Xu C, Zhang YF, Zhang Z, Wang L, Bian J. TimeRAF: retrieval-augmented foundation model for zero-shot time series forecasting. IEEE Trans Know Data Eng. 2025;37(9):5654–65.
61. Lyu X, Han Y, Wang W, Qian H, Tsang I, Zhang X. Cross-context backdoor attacks against graph prompt learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 2094–105. doi:10.1145/3637528.3671956.
62. Wang D, Li B, Song B, Chen C, Yu FR. HSMH: a hierarchical sequence multi-hop reasoning model with reinforcement learning. IEEE Trans Know Data Eng. 2024;36(4):1638–49. doi:10.1109/TKDE.2023.3303617.
63. Aydın A, Arslan A, Dincer BT. A set of novel HTML document quality features for Web information retrieval: including applications to learning to rank for information retrieval. Expert Syst Appl. 2024;246(81):123177. doi:10.1016/j.eswa.2024.123177.
64. Hu Z, Chen Y, Zhao M, Li R, Wang L. UniRQR: a unified model for retrieval decision, query, and response generation in internet-based knowledge dialogue systems. Expert Syst Appl. 2025;270:126494. doi:10.1016/j.eswa.2025.126494.
65. Abaho M, Alfaifi YH. Select and augment: enhanced dense retrieval knowledge graph augmentation. J Artif Intell Res. 2023;78:269–85. doi:10.1613/jair.1.14365.

66. Pan M, Zhou S, Chen J, Huang EA, Huang JX. A semantic framework for enhancing pseudo-relevance feedback with soft negative sampling and contrastive learning. Inform Process Manag. 2025;62(3):104058. doi:10.1016/j.ipm.2024.104058.
67. Nguyen C, Nguyen P, Nguyen LM. Retrieve–revise–refine: a novel framework for retrieval of concise entailing legal article set. Inform Process Manag. 2025;62(1):103949. doi:10.1016/j.ipm.2024.103949.
68. Guo C, Tian Z, Tang J, Li S, Wang T. Multi-pattern retrieval-augmented framework for Text-to-SQL with Poincaré-Skeleton retrieval and meta-instruction reasoning. Inform Process Manag. 2025;62(3):103978. doi:10.1016/j.ipm.2024.103978.
69. Wang Z, Long C, Cong G. Similar sports play retrieval with deep reinforcement learning. IEEE Trans Know Data Eng. 2023;35(4):4253–66. doi:10.1109/TKDE.2021.3136881.
70. Chen Y, Fang X, Liu Y, Zheng W, Kang P, Han N, et al. Two-step strategy for domain adaptation retrieval. IEEE Trans Know Data Eng. 2024;36(2):897–912. doi:10.1109/tkde.2023.3289882.
71. Zhou Y, Yao J, Wu L, Dou Z, Wen JR. WebUltron: an ultimate retriever on webpages under the model-centric paradigm. IEEE Trans Know Data Eng. 2024;36(9):4996–5006. doi:10.1109/TKDE.2023.3332858.
72. Tan H, Zhan S, Lin H, Zheng HT, Chan WK. QAEA-DR: a unified text augmentation framework for dense retrieval. IEEE Trans Know Data Eng. 2025;37(6):3669–83. doi:10.1109/TKDE.2025.3543203.
73. Salemi A, Kallumadi S, Zamani H. Optimization methods for personalizing large language models through retrieval augmentation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24; 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 752–62. doi:10.1145/3626772.3657783.
74. Faggioli G, Ferro N, Perego R, Tonellotto N. Dimension importance estimation for dense information retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24; 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 1318–28. doi:10.1145/3626772.3657691.
75. Zhang W, Li Y, Li Z, Sun H, Gao X, Liu X. ModelGalaxy: a versatile model retrieval platform. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24; 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 2771–5. doi:10.1145/3626772.3657676.
76. Zeng H, Kallumadi S, Alibadi Z, Nogueira R, Zamani H. A personalized dense retrieval framework for unified information access. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23; 2023 Jul 23–27; Taipei, Taiwan. New York, NY, USA: Association for Computing Machinery; 2023. p. 121–30. doi:10.1145/3539618.3591626.
77. Jiang W, Zhang S, Han B, Wang J, Wang B, Kraska T. PipeRAG: fast retrieval-augmented generation via adaptive pipeline parallelism. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, KDD '25; 2025 Aug 3–7; Toronto, ON, Canada. New York, NY, USA: Association for Computing Machinery; 2025. p. 589–600. doi:10.1145/3690624.3709194.
78. Li Z, Li X, Tao C, Feng J, Shen T, Xu C, et al. RetriEVAL: evaluating text generation with contextualized lexical match. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25; 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 934–43. doi:10.1145/3701551.3703581.
79. Wen L, Wang Y, Zhang D, Chen G. Visual matching is enough for scene text retrieval. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM '23; 2023 Feb 27–Mar 3; Singapore. New York, NY, USA: Association for Computing Machinery; 2023. p. 447–55. doi:10.1145/3539597.3570428.
80. Zhu C, Hu X, Wu H, Qin C, Zhu H, Xiong H. Enhancing job recommendations with LLM-based resume completion: a behavior-denoised alignment approach. Inform Process Manag. 2025;62(6):104261. doi:10.1016/j.ipm.2025.104261.

81. Sachdeva N, Coleman B, Kang WC, Ni J, Caverlee J, Hong L, et al. Improving data efficiency for recommenders and LLMs. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 790–2. doi:10.1145/3640457.3688052.
82. Wang J, Lu H, Liu Y, Ma H, Wang Y, Gu Y, et al. LLMs for user interest exploration in large-scale recommendation systems. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 872–7. doi:10.1145/3640457.3688161.
83. Wan Z, Yin B, Xie J, Jiang F, Li X, Lin W. LARR: large language model aided real-time scene recommendation with semantic understanding. In: Proceedings of the 18th ACM Conference on Recommender Systems, (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 23–32. doi:10.1145/3640457.3688135.
84. Cui Y, Liu F, Wang P, Wang B, Tang H, Wan Y, et al. Distillation matters: empowering sequential recommenders to match the performance of large language models. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 507–17. doi:10.1145/3640457.3688118.
85. Zhang X, Li Y, Wang J, Sun B, Ma W, Sun P, et al. Large language models as evaluators for recommendation explanations. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 33–42. doi:10.1145/3640457.3688075.
86. Xi Y, Liu W, Lin J, Cai X, Zhu H, Zhu J, et al. Towards open-world recommendation with knowledge augmentation from large language models. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 12–22. doi:10.1145/3640457.3688104.
87. Yang T, Chen L. Unleashing the retrieval potential of large language models in conversational recommender systems. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 43–52. doi:10.1145/3640457.3688146.
88. Petruzzelli A, Musto C, Di Carlo MC, Tempesta G, Semeraro G. Recommending healthy and sustainable meals exploiting food retrieval and large language models. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 1057–61. doi:10.1145/3640457.3688193.
89. Petruzzelli A, Musto C, Laraspata L, Rinaldi I, de Gemmis M, Lops P, et al. Instructing and prompting large language models for explainable cross-domain recommendations. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 298–308. doi:10.1145/3640457.3688137.
90. Kong X, Wu J, Zhang A, Sheng L, Lin H, Wang X, et al. Customizing language models with instance-wise LoRA for sequential recommendation. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al., editors. Proceedings of the 38th International Conference on Neural Information Processing System; 2024 Dec 10–15; Vancouver, BC, Canada. Red Hook, NY, USA: Curran Associates, Inc.; 2024. Vol. 37, p. 113072–95.
91. Liu R, Chen H, Bei Y, Shen Q, Zhong F, Wang S, et al. Fine tuning out-of-vocabulary item recommendation with user sequence imagination. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al., editors. Proceedings of the 38th International Conference on Neural Information Processing Systems; 2024 Dec 10–15; Vancouver, BC, Canada. Red Hook, NY, USA: Curran Associates, Inc.; 2024. p. 8930–55.
92. Na H, Gang M, Ko Y, Seol J, Lee Sg. Enhancing large language model based sequential recommender systems with pseudo labels reconstruction. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov 12–16; Miami, FL, USA. p. 7213–22. doi:10.18653/v1/2024.findings-emnlp.423.

93. Ma Q, Ren X, Huang C. XRec: large language models for explainable recommendation. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov 12–16; Miami, FL, USA. p. 391–402. doi:10.18653/v1/2024.findings-emnlp.22.
94. Zhang H, Zhu Q, Dou Z. Enhancing reranking for recommendation with LLMs through user preference retrieval. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics; 2025 Jan 19–24; Abu Dhabi, UAE. p. 658–71.
95. Cao Y, Mehta N, Yi X, Hulikal Keshavan R, Heldt L, Hong L, et al. Aligning large language models with recommendation knowledge. In: Duh K, Gomez H, Bethard S, editors. Findings of the Association for Computational Linguistics: NAACL 2024; 2024 Jun 16–21; Mexico City, Mexico. p. 1051–66.
96. Li C, Deng Y, Hu H, Kan MY, Li H. ChatCRS: incorporating external knowledge and goal guidance for LLM-based conversational recommender systems. In: Chiruzzo L, Ritter A, Wang L, editors. Findings of the Association for Computational Linguistics: NAACL 2025; 2025 Apr 29–May 4; Albuquerque, NM, USA. p.295–312. doi:10.18653/v1/2025.findings-naacl.17.
97. Bismay M, Dong X, Caverlee J. ReasoningRec: bridging personalized recommendations and human-interpretable explanations through LLM reasoning. In: Chiruzzo L, Ritter A, Wang L, editors. Findings of the Association for Computational Linguistics: NAACL 2025; 2025 Apr 29–May 4; Albuquerque, NM, USA. p. 8132–48. doi:10.18653/v1/2025.findings-naacl.454.
98. Wang Y, Jiang Z, Chen Z, Yang F, Zhou Y, Cho E, et al. RecMind: large language model powered agent for recommendation. In: Duh K, Gomez H, Bethard S, editors. Findings of the Association for Computational Linguistics: NAACL 2024; 2024 Jun 16–21; Mexico City, Mexico. p. 4351–64. doi:10.18653/v1/2024.findings-naacl.271.
99. Shi W, He X, Zhang Y, Gao C, Li X, Zhang J, et al. Large language models are learnable planners for long-term recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24); 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 1893–903. doi:10.1145/3626772.3657683.
100. Xu S, Hua W, Zhang Y. OpenP5: an open-source platform for developing, training, and evaluating LLM-based recommender systems. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24); 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 386–94. doi:10.1145/3626772.3657883.
101. Liao J, Li S, Yang Z, Wu J, Yuan Y, Wang X, et al. LLaRA: large language-recommendation assistant. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). New York, NY, USA: Association for Computing Machinery; 2024. p. 1785–95. doi:10.1145/3626772.3657690.
102. Wang J, Karatzoglou A, Arapakis I, Jose JM. Reinforcement learning-based recommender systems with large language models for state reward and action modeling. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). New York, NY, USA: Association for Computing Machinery; 2024. p. 375–85. doi:10.1145/3626772.3657767.
103. Li P, de Rijke M, Xue H, Ao S, Song Y, Salim FD. Large language models for next point-of-interest recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). New York, NY, USA: Association for Computing Machinery; 2024. p. 1463–72. doi:10.1145/3626772.3657840.
104. Sharma A, Li H, Li X, Jiao J. Optimizing novelty of top-k recommendations using large language models and reinforcement learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 5669–79. doi:10.1145/3637528.3671618.
105. Wei W, Ren X, Tang J, Wang Q, Su L, Cheng S, et al. LLMRec: large language models with graph augmentation for recommendation. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24); 2024 Mar 4–8; Merida, Mexico. New York, NY, USA: Association for Computing Machinery; 2024. p. 806–15. doi:10.1145/3616855.3635853.

106. Shi J, Zhao J, Wu X, Xu R, Jiang YH, He L. Mitigating reasoning hallucination through multi-agent collaborative filtering. Expert Syst Appl. 2025;263:125723. doi:10.1016/j.eswa.2024.125723.
107. Ren R, Ma J, Zheng Z. Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method. Expert Syst Appl. 2025;278(2):127330. doi:10.1016/j.eswa.2025.127330.
108. Walek B, Müller P. A text-based recommender system for recommending relevant news articles. Expert Syst Appl. 2025;266(6):125816. doi:10.1016/j.eswa.2024.125816.
109. Yu D, Yu T, Wang D, Wang S. Long tail service recommendation based on cross-view and contrastive learning. Expert Syst Appl. 2024;238(1):121957. doi:10.1016/j.eswa.2023.121957.
110. Wang Z, Zeng J, Wen J, Gao M, Zhou W. Point-of-interest recommendation using deep semantic model. Expert Syst Appl. 2023;231(8):120727. doi:10.1016/j.eswa.2023.120727.
111. D'Asaro F, De Luca S, Bongiovanni L, Rizzo G, Papadopoulos S, Schinas M, et al. Zero-shot content-based crossmodal recommendation system. Expert Syst Appl. 2024;258(4):125108. doi:10.1016/j.eswa.2024.125108.
112. Bauer J, Jannach D. Hybrid session-aware recommendation with feature-based models. User Model User Adapt Interact. 2024;34(3):691–728. doi:10.1007/s11257-023-09379-6.
113. Bevec M, Tkalčič M, Pesek M. Hybrid music recommendation with graph neural networks. User Model User Adapt Interact. 2024;34(5):1891–928. doi:10.1007/s11257-024-09410-4.
114. Gheewala S, Xu S, Yeom S. Deep shared learning and attentive domain mapping for cross-domain recommendation. User Model User Adapt Interact. 2024;34(5):1981–2038. doi:10.1007/s11257-024-09416-y.
115. Ghadami A, Tran T. TriDeepRec: a hybrid deep learning approach to content-and behavior-based recommendation systems. User Model User Adapt Interact. 2024;34(5):2085–114. doi:10.1007/s11257-024-09418-w.
116. Sun W, Ma M, Ren P, Lin Y, Chen Z, Ren Z, et al. Parallel split-join networks for shared account cross-domain sequential recommendations. IEEE Trans Know Data Eng. 2023;35(4):4106–23. doi:10.1109/tkde.2021.3130927.
117. Göpfert C, Haig A, Hsu CW, Chow Y, Vendrov I, Lu T, et al. Discovering personalized semantics for soft attributes in recommender systems using concept activation vectors. ACM Trans Recomm Syst. 2024;2(4):1–37. doi:10.1145/3658675.
118. Cavenaghi E, Zanga A, Stella F, Zanker M. Towards a causal decision-making framework for recommender systems. ACM Trans Recomm Syst. 2024;2(2):1–34. doi:10.1145/3629169.
119. Li M, Ariannezhad M, Yates A, De Rijke M. Who will purchase this item next? reverse next period recommendation in grocery shopping. ACM Trans Recomm Syst. 2023;1(2):1–32. doi:10.1145/3595384.
120. Srba I, Moro R, Tomlein M, Pecher B, Simko J, Stefancova E, et al. Auditing youtube's recommendation algorithm for misinformation filter bubbles. ACM Trans Recomm Syst. 2023;1(1):1–33. doi:10.1145/3568392.
121. Yu Y, Sugiyama K, Jatowt A. Beyond recommendations: sequential recommendation with collaborative explanation. ACM Trans Recomm Syst. 2026;4(3):1–31. doi:10.1145/3731458.
122. Liu D, Greene D, Li I, Jiang X, Dong R. Topic-centric explanations for news recommendation. ACM Trans Recomm Syst. 2025;3(2):1–25. doi:10.1145/3680295.
123. Chiang HY, Chen YS, Song YZ, Shuai HH, Chang JS. Shilling black-box review-based recommender systems through fake review generation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23; 2023 Aug 6–13; Long Beach, CA, USA. New York, NY, USA: Association for Computing Machinery; 2023. p. 286–97. doi:10.1145/3580305.3599502.
124. Tian Z, Ou Z, Zhu Y, Lyu S, Zhang H, Xiao J, et al. Multi-SEA: multi-stage semantic enhancement and aggregation for image-text retrieval. Inform Process Manag. 2025;62(5):104165. doi:10.1016/j.ipm.2025.104165.
125. Jing P, Sun H, Nie L, Li Y, Su Y. Deep multi-modal hashing with semantic enhancement for multi-label micro-video retrieval. IEEE Trans Know Data Eng. 2024;36(10):5080–91. doi:10.1109/tkde.2023.3337077.
126. Li Y, Du J, Wang C, Liu Z, Zhu X, Lin C. CROSS: feedback-oriented multi-modal dynamic alignment in recommendation systems. ACM Trans Recomm Syst. 2026;4(3):1–24. doi:10.1145/3734527.
127. Yi Z, Long Z, Ounis I, Macdonald C, McCreadie R. Enhancing recommender systems: deep modality alignment with large multi-modal encoders. ACM Trans Recomm Syst. 2025;3(4):1–25. doi:10.1145/3718099.
128. Zeng J, Yu Y, Wen J, Jiang W, Cheng L. Personalized dynamic attention multi-task learning model for document retrieval and query generation. Expert Syst Appl. 2023;213(6):119026. doi:10.1016/j.eswa.2022.119026.

129. Sun S, Zhang K, Li J, Yu M, Hou K, Wang Y, et al. Retriever-generator-verification: a novel approach to enhancing factual coherence in open-domain question answering. Inform Process Manag. 2025;62(4):104147. doi:10.1016/j.ipm.2025.104147.
130. Li Y, Yang N, Wang L, Wei F, Li W. Generative retrieval for conversational question answering. Inform Process Manag. 2023;60(5):103475. doi:10.1016/j.ipm.2023.103475.
131. Yang Z, Xiang J, You J, Li Q, Liu W. Event-oriented visual question answering: the E-VQA dataset and benchmark. IEEE Trans Know Data Eng. 2023;35(10):10210–23. doi:10.1109/TKDE.2023.3267036.
132. Yuan Q, Yuan Y, Wen Z, Wang H, Tang S. An effective framework for enhancing query answering in a heterogeneous data lake. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23); 2023 Jul 23–27; Taipei, Taiwan. New York, NY, USA: Association for Computing Machinery; 2023. p. 770–80. doi:10.1145/3539618.3591637.
133. Kusano G. Data augmentation using reverse prompt for cost-efficient cold-start recommendation. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 861–5. doi:10.1145/3640457.3688159.
134. Li H, Zhang C, Jia X, Gao Y, Chen C. Adaptive label correlation based asymmetric discrete hashing for cross-modal retrieval. IEEE Trans Know Data Eng. 2023;35(2):1185–99. doi:10.1109/tkde.2021.3102119.
135. He S, Wang W, Wang Z, Xu X, Yang Y, Wang X, et al. Category alignment adversarial learning for cross-modal retrieval. IEEE Trans Know Data Eng. 2023;35(5):4527–38. doi:10.1109/tkde.2022.3153962.
136. Zhang C, Li H, Gao Y, Chen C. Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval. IEEE Trans Know Data Eng. 2023;35(6):6475–88. doi:10.1109/tkde.2022.3172216.
137. Huang J, Kang P, Han N, Chen Y, Fang X, Gao H, et al. Two-stage asymmetric similarity preserving hashing for cross-modal retrieval. IEEE Trans Know Data Eng. 2024;36(1):429–44. doi:10.1109/tkde.2023.3283984/mm1.
138. Hostnik M, Robnik-Šikonja M. Retrieval-augmented code completion for local projects using large language models. Expert Syst Appl. 2025;292:128596. doi:10.1016/j.eswa.2025.128596.
139. Ren T, Zhang Z, Jia B, Zhang S. Retrieval-augmented generation-aided causal identification of aviation accidents: a large language model methodology. Expert Syst Appl. 2025;278:127306. doi:10.1016/j.eswa.2025.127306.
140. Aluri GS, Sharma S, Sharma T, Delgado J. Playlist search reinvented: LLMs behind the curtain. In: Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24); 2024 Oct 14–18; Bari, Italy. New York, NY, USA: Association for Computing Machinery; 2024. p. 813–5. doi:10.1145/3640457.3688047.
141. Salemi A, Zamani H. Towards a search engine for machines: unified ranking for multiple retrieval-augmented large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24); 2024 Jul 14–18; Washington, DC, USA. New York, NY, USA: Association for Computing Machinery; 2024. p. 741–51. doi:10.1145/3626772.3657733.
142. Che TY, Mao XL, Lan T, Huang H. A hierarchical context augmentation method to improve retrieval-augmented LLMs on scientific papers. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 243–54. doi:10.1145/3637528.3671847.
143. Sheng Y, Gandhe S, Kanagal B, Edmonds N, Fisher Z, Tata S, et al. Measuring an LLM's proficiency at using APIs: a query generation strategy. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 5680–9. doi:10.1145/3637528.3671592.
144. Chen L, Xu F, Li N, Han Z, Wang M, Li Y, et al. Large language model-driven meta-structure discovery in heterogeneous information network. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 307–18. doi:10.1145/3637528.3671965.
145. Hu Z, Wang C, Shu Y, Paik HY, Zhu L. Prompt perturbation in retrieval-augmented generation based large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24); 2024 Aug 25–29; Barcelona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 1119–30. doi:10.1145/3637528.3671932.

146. Wang Y, Yu J, Yao Z, Zhang J, Xie Y, Tu S, et al. SoAy: a solution-based LLM API-using methodology for academic information seeking. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25); 2025 Aug 3–7; Toronto, ON, Canada. New York, NY, USA: Association for Computing Machinery; 2025. p. 2660–71. doi:10.1145/3690624.3709412.
147. Amirizaniani M, Sivachenko M, Lavergne A, Shah C, Mashhadi A. How does memorization impact LLMs' social reasoning? An assessment using seen and unseen queries. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25); 2025 Mar 10–14; Hannover, Germany. New York, NY, USA: Association for Computing Machinery; 2025. p. 924–33. doi:10.1145/3701551.3703576.
148. Rahdari B, Ding H, Fan Z, Ma Y, Chen Z, Deoras A, et al. Logic-scaffolding: personalized aspect-instructed recommendation explanation generation using LLMs. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24); 2024 Mar 4–8; Merida, Mexico. New York, NY, USA: Association for Computing Machinery; 2024. p. 1078–81. doi:10.1145/3616855.3635689.
149. Liu P. Unsupervised corrupt data detection for text training. Expert Syst Appl. 2024;248(2):123335. doi:10.1016/j.eswa.2024.123335.
150. Arazzi M, Marconi Sciarroni M, Nocera A, Storti E. RAG-IoE: IoT context-aware information retrieval with large language models in industry 5.0. ACM Trans Internet Things. 2025;6(4):1–31.
151. Mukherjee K. An adaptive RAG-based question-answering system in the context of industry 5.0. J Comput Anal Appl. 2025;34(7):259–74.
152. Zhou T, Wan Y, Liu Y, Kumar M. Enabling interactive AI in industry 5.0 with RAG-enhanced GenAI chatbots. In: Proceedings of the 2025 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC); 2025 Jun 16–19; Valencia, Spain. p. 1–10.
153. Ponnock J, Kenneally G, Briggs MR, Yeo E, Patterson T III, Kinberg N, et al. Real-time RAG for the identification of supply chain vulnerabilities. arXiv:2509.10469. 2025.
154. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2020. p. 9459–74.
155. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 6769–81. doi:10.18653/v1/2020.emnlp-main.550.
156. Izacard G, Grave E. Leveraging passage retrieval with generative models for open domain question answering. In: Merlo P, Tiedemann J, Tsarfaty R, editors. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 2021 Apr 19–23; Online. p. 874–80. doi:10.18653/v1/2021.eacl-main.74.
157. Larson J, Truitt S. GraphRAG: unlocking LLM discovery on narrative private data. Micro Res Blog. 2024.
158. Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, et al. From local to global: a graph rag approach to query-focused summarization. arXiv:2404.16130. 2024.
159. Contal E, McGoldrick G. Ragsys: item-cold-start recommender as rag system. arXiv:2405.17587. 2024.
160. Wang S, Fan W, Feng Y, Shanru L, Ma X, Wang S, et al. Knowledge graph retrieval-augmented generation for LLM-based recommendation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2025 Jul 27–Aug 1; Vienna, Austria. p. 27152–68.
161. Raja R, Vats A, Vats A, Majumder A. A comprehensive review on harnessing large language models to overcome recommender system challenges. arXiv:2507.21117. 2025.